



IMPERIAL INSTITUTE  
OF  
AGRICULTURAL RESEARCH, PUSA.

THE ANNALS  
of  
MATHEMATICAL  
STATISTICS

Published and Lithoprinted by  
EDWARDS BROTHERS, INC.  
ANN ARBOR, MICH.



THE ANNALS OF MATHEMATICAL STATISTICS is  
affiliated with the American Statistical Association and is devoted  
to the theory and application of Mathematical Statistics.

Published Quarterly: March, June, September, December

*Four Dollars per annum*

H. C. CARVER, *Editor*

A. L. O'TOOLE, *Associate Editor*

*The Annals is not copyrighted: any articles or tables appearing therein may  
be reproduced in whole or in part at any time if accompanied by  
the proper reference to this publication*

Address: ANNALS OF MATHEMATICAL STATISTICS  
Post Office Box 171, Ann Arbor, Michigan

# CONTENTS OF VOLUME V

The Accuracy of Computation with Approximate Numbers	1
<i>Helen M. Walker and Vera Sanford</i>	
Combining Two Probability Functions . . . . .	13
<i>William Dowell Baten</i>	
On the Systematic Fitting of Straight Line Trends by Stencil and Calculating Machine . . . . .	21
<i>Herbert A. Toops</i>	
Statistical Analysis of One-Dimensional Distributions . .	30
<i>Robert Schmidt, Kiel, Germany</i>	
<i>Editorial: A New Type of Average for Security Prices . .</i>	<i>73</i>
<i>H. C. Carver</i>	
On a New Method of Computing Non-Linear Regression Curves . . . . .	81
<i>Walter Andersson, Lund, Sweden</i>	
The Standard Error of Any Analytic Function of a Set of Parameters Evaluated by the Method of Least Squares	107
<i>Walter A. Hendricks</i>	
Transformation of Non-Normal Frequency Distributions into Normal Distributions . . . . .	113
<i>G. A. Baker</i>	
Invariants and Covariants of Certain Frequency Curves . .	124
<i>Richmond T. Zoch</i>	
Quadrature of the Normal Curve . . . . .	136
<i>E. R. Enlow</i>	
<i>Editorial: On a Best Value of <math>R</math> in Sample of <math>R</math> from a Finite Population of <math>N</math> . . . . .</i>	<i>146</i>
<i>A. L. O'Toole</i>	
<i>Editorial: Punched Card Systems and Statistics . . . . .</i>	<i>153</i>
<i>H. C. Carver</i>	

## CONTENTS OF VOLUME V—Continued

The Method of Path Coefficients . . . . .	161
<i>Sewall Wright</i>	

Mathematical Foundation for a Method of Statistical Analysis of Household Budgets . . . . .	216
<i>John W. Boldyreff</i>	

On the Relative Stability of the Median and Arithmetic Mean, with Particular Reference to Certain Frequency Distributions which can be Dissected into Normal Distributions .	227
<i>Harry S. Pollard</i>	

An Application of Characteristic Functions to the Distribution Problem of Statistics . . . . .	263
<i>Solomon Kullback</i>	

On Measures of Contingency . . . . .	308
<i>Frank M. Weida</i>	

Note on Koshal's Method of Improving the Parameters of Curves by the Use of the Method of Maximum Likelihood . . . . .	320
<i>R. J. Myers</i>	

The Adequacy of "Student's" Criterion of Deviations in Small Sample Means . . . . .	324
<i>Alan E. Treloar and Marian A. Wilder</i>	

# THE ACCURACY OF COMPUTATION WITH APPROXIMATE NUMBERS

By

HELEN M. WALKER

*Teachers College, Columbia University*

and

VERA SANFORD

*Oneonta State Normal School.*

## 1. *General Considerations.*

The number of figures necessarily free from error in the result of a piece of computation may be determined by studying the relation between the number of digits in the result and the number of digits in the maximum error of the computation. It is the purpose of this essay to derive rules for the determination of the number of digits which are certain to be correct in computations based on measurement, but it must be understood that these rules state the minimum number of correct digits so that the result of a specific piece of computation may be accurate for more places than the rules indicate.

## 2. *Notation.*

Since the location of the decimal point has no connection with significant figures in a given number, it is assumed that the decimal point follows the last significant figure in each of the original numbers, the argument being somewhat simplified by this assumption. Accordingly the greatest error in the statement of the original numbers is  $\pm 0.5$ . Let  $A$  and  $B$  be the true values of two numbers such that  $A = a \cdot 10^m$  and  $B = b \cdot 10^n$ , where  $m$  and  $n$  are positive integers and where  $0.1 \leq a < 1.0$  and where  $0.1 \leq b < 1.0$ . Then by the convention adopted above, the number of significant figures in  $A$  and  $B$  are  $m$  and  $n$  respectively, and the observed values are not less than  $A - 0.5$  and  $B - 0.5$  and not more than

$A + 0.5$  and  $B + 0.5$ . Let  $\epsilon$  represent the maximum error in the computation and let  $\epsilon'$  be the value of the largest term in the expansion of  $\epsilon$ .

### 3. Products.

The greatest error in the product of  $A$  and  $B$  will occur when each is in excess by 0.5, the value of this error being

$$\epsilon = (A+0.5)(B+0.5) - AB = \frac{1}{2}a \cdot 10^m + \frac{1}{2}b \cdot 10^n + .25$$

For  $ab \geq 0.1$ ,  $AB$  has  $m+n$  digits to the left of the decimal point.

For  $ab < 0.1$ ,  $AB$  has  $m+n-1$  digits to the left of the decimal point. The cases to be considered are

$$(I) \quad m = n = 1$$

$$(II) \quad m = n > 1$$

$$(III) \quad m - n = 1$$

$$(IV) \quad m - n > 1$$

(I) Let  $m = n = 1$ . In this extreme case each factor consists of a single digit and the product consists of one or two digits. In this case, the figure in the unit's place is always affected by the maximum error and the figure in the ten's place, when present, is generally so affected.

$$(II) \quad \text{Let } m = n > 1. \text{ Here } AB = ab \cdot 10^{2n} \text{ and}$$

$$\epsilon = \frac{1}{2}(a+b) \cdot 10^n + .25.$$

$$\text{But } 0.1 \leq \frac{a+b}{2} < 1.0$$

$$\text{and therefore } 10^{n-1} + \frac{1}{4} \leq \epsilon < 10^n + \frac{1}{4}.$$

The following conditions are possible:

Either (1)  $ab \geq 0.1$  and  $AB$  has  $2n$  digits to the left of the decimal point,

or (2)  $ab < 0.1$  and  $AB$  has  $2n-1$  digits to the left of the decimal point.

Either (3)  $\epsilon$  has  $n$  digits to the left of the point,

or (4)  $\epsilon$  has  $n+1$  digits to the left of the point, the first one being 1 and all the others to the left of the decimal point being zero. This can occur only

when  $\epsilon$  is very near its maximum value. For example, when  $n=4$ , the value of  $\epsilon$  must be less than 10000.25.

Then if conditions (1) and (3) are fulfilled, the result has  $n$  more digits to the left of the decimal point than has the error. A subsequent proof will show that this means that at least  $n-1$  places in the result are free from error. Under conditions (1) and (4), the difference in the number of digits is  $n-1$ , and at least  $n-2$  are not affected by the error. Similarly under conditions (2) and (3),  $n-2$  digits are not affected by the error. Conditions (2) and (4) cannot occur simultaneously.

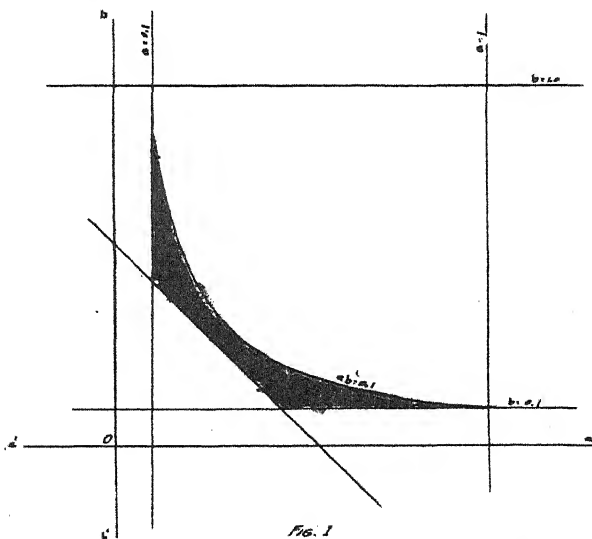


Fig. 1

The proof that conditions (2) and (4) are incompatible with the conditions that  $0.1 \leq a < 1.0$  and  $0.1 \leq b < 1.0$  may be obtained from fig. 1. The area within which these limits hold for  $a$  and for  $b$  is the area of the square bounded by  $a=1.0$ ,  $a=0.1$ ,  $b=1.0$ ,  $b=0.1$ , the numerical value of this area being 0.81. The region within which  $ab < 0.1$  is the area below the hyperbola  $ab=0.1$ . The region within which  $\frac{a+b}{2}$  is larger than a specified value is

the region above the line  $a+b=k$ . Therefore the probability that all of these conditions shall be simultaneously fulfilled is the ratio of the shaded region in fig. 1 to the total area of the square, or to 0.81. When  $\frac{a+b}{2} > 5$ , this probability is only 0.000,014,8 and when  $\frac{a+b}{2} > 0.55$ , the probability vanishes altogether.

(III) Let  $m-n=1$ . Then  $\epsilon = \frac{1}{2}(10a+b) \cdot 10^n + .25$ .

But  $1.1 \leq 10a+b < 11$ .

Now either  $n=1$  or  $n>1$ .

Let  $n=1$ . Then  $5.75 \leq \epsilon < 55.25$ .

Either (1)  $ab \geq 0.1$  and  $AB$  has 3 digits to the left of the decimal point,

or (2)  $ab < 0.1$  and  $AB$  has 2 digits to the left of the decimal point.

Either (3)  $10a+b < 1.95$  and  $5.75 \leq \epsilon < 10$ ,

or (4)  $10a+b \geq 1.95$  and  $10 \leq \epsilon < 55.25$ .

If conditions (1) and (3) are met, the product has 2 more digits to the left of the decimal point than has the error. Thus one or two places will in general be free from error. Under conditions (1) and (4) or conditions (2) and (3) the number of such places free from error is 0 or 1. By an analysis similar to that given under (II) it appears that there is about one chance in ten that conditions (2) and (4) should be simultaneously met, in which case no place would be free from error.

Let  $n=2$ . Then  $55.25 < \epsilon < 550.25$ .

Either (1)  $ab \geq 0.1$  and  $AB$  has 5 digits to the left of the decimal point,

or (2)  $ab < 0.1$  and  $AB$  has 4 digits to the left of the decimal point.

Either (3)  $10a+b < 1.995$  and  $55.25 \leq \epsilon < 100$ .

or (4)  $10a+b \geq 1.995$  and  $100 \leq \epsilon < 550.25$ .

If conditions (1) and (3) are met, the product has either 2 or 3 places free from error. Under conditions (2) and (3) or

conditions (1) and (4), the product has 1 or 2 places free from error. Simultaneous fulfilment of conditions (2) and (4) will be rare but not impossible. However in this case, the first digit of the error cannot be larger than 5, hence, as shown later, the number of digits free from error in the result will usually be the number in the product minus the number in the error, rather than one less than that.

For  $n > 2$ , the constant 0.25 forms a still smaller proportion of the error. Hence for larger values of  $n$ , if  $m - n = 1$ , the product may be expected to have  $n$  or  $n-1$  places free from error.

(IV) Let  $m - n > 1$ . Here  $m \geq 3$  and therefore the terms  $\frac{1}{2} b \cdot 10^m$  and 0.25 are negligible in comparison with  $\frac{1}{2} a \cdot 10^m$  and may be disregarded, since neither of them can affect the first place in the error. Then  $\epsilon' = \frac{1}{2} a \cdot 10^m$ , and therefore

$$0.5 (10^{m-1}) < \epsilon' < 0.5 (10^m).$$

Either (1)  $a b \geq 0.1$  and  $AB$  has  $m+n$  places to the left of the decimal point,

or (2)  $a b < 0.1$  and  $AB$  has  $m+n-1$  places to the left of the decimal point.

Either (3)  $a < 2$  and  $\epsilon' < 10^{m-1}$  so that  $\epsilon$  has  $m-1$  places to the left of the decimal point,

or (4)  $a \geq 2$  and  $10^{m-1} < \epsilon' < 0.5 (10^m)$ , so that  $\epsilon$  has  $m$  places to the left of the decimal point.

Conditions (1) and (3) would leave either  $n+1$  or  $n$  places free from error in the product. Conditions (2) and (3) or conditions (1) and (4) would leave either  $n$  or  $n-1$  places free from error. If conditions (2) and (4) are met, there would be  $n-1$  places in the product to the left of the first digit in the error, and since this first digit is not more than 5, the error is not likely to affect the preceding digit. See section 6.

*In general, therefore, if there are  $n$  significant figures in the less accurate of two approximations, the product of the two approximations will have  $n$  or  $n-1$  digits free from error. The product of*



two such numbers should be rounded off until it contains only as many significant figures as the less accurate of the two numbers. The last digit in the product may then contain some error.

#### 4. Quotients.

The greatest error which can occur in a quotient arises when the dividend is in excess by 0.5 and the divisor in defect by the same amount.

$$\text{Then } \epsilon = \frac{A + .5}{B - .5} - \frac{A}{B} = \frac{b + a \cdot 10^{m-n}}{b(2b \cdot 10^n - 1)}$$

We must consider separately the cases (I)  $m = n$

(II)  $m > n$

(III)  $m < n$

(I) Let  $m = n$ . Then

$$\epsilon = \frac{b+a}{b(2b \cdot 10^n - 1)}$$

(a) If  $m = n = 1$  there are 81 possible quotients of one-place numbers, and an examination of these shows that in only thirty-one of these cases the first digit is free from error.

(b) The case for  $m = n > 1$  should be studied for specific values of  $n$ . In general, however, the error is large as  $a \rightarrow 1$  and  $b \rightarrow 0.1$ , and is small as  $a \rightarrow 0.1$  and  $b \rightarrow 1.0$ . Also as  $n$  increases, the influence of the constant term in the denominator becomes less. Therefore in general

$$\frac{0.55}{10} < \frac{1.1}{2 \cdot 10^{n-1}} < \epsilon < \frac{1.1}{0.1[0.2(10^n) - 1]} = \frac{55}{10^n - 5} < \frac{0.61}{10^{n-2}}$$

If  $a \rightarrow 1.0$ ,  $b \rightarrow 0.1$  and  $n$  is large, then  $\epsilon < \frac{0.61}{10^{n-2}}$  and the error has at least  $n-2$  zeros to the left of the first digit. In this case,  $\frac{A}{B}$  has one digit to the left of the decimal point, so that the quotient will have at least  $n-1$  digits to the left of the first digit in the error.

If  $a \rightarrow 0.1$ ,  $b \rightarrow 1.0$ , and  $n$  is large, then  $\epsilon > \frac{0.55}{10^n}$ , and the error has  $n$  zeros to the left of the first digit. In this case  $\frac{A}{B}$  has no digits to the left of the decimal, so that the quotient will again have  $n-1$  digits to the left of the first digit in the error.

Furthermore,  $\epsilon = \left(\frac{a}{b} + 1\right)(2b \cdot 10^n - 1)$ , OR

$$\epsilon = \frac{1}{2b \cdot 10^n} + \frac{a}{2b^2 \cdot 10^n} + \text{higher powers of } 10^{-n}.$$

Then if  $a \geq b$ ,  $\frac{1}{b \cdot 10^n} \leq \epsilon' < \frac{55}{b \cdot 10^n}$ . We then have 1 digit to the left of the decimal point in the quotient, and either  $n-1$  or  $n-2$  zeros preceding the first digit in the error.

$$\text{If } a < b, \quad \frac{0.55}{b \cdot 10^n} < \epsilon' < \frac{1}{b \cdot 10^n}$$

$$\text{since } 0.1 > \frac{a}{b} > 10.$$

In this case we have no digits to the left of the decimal point in the quotient, and either  $n$  or  $n-1$  zeros preceding the first digit in the quotient.

(II) Let  $m > n$ .

(a) Let  $m-n=1$ .

$$\text{Then } \epsilon = \frac{b+10a}{b(2b \cdot 10^{n-1}-1)} = \left(1 + \frac{10a}{b}\right) \left\{ (2b \cdot 10^n)^{-1} + (2b \cdot 10^n)^{-2} + \dots \right\}.$$

$$\therefore \epsilon = \left(1 + \frac{10a}{b}\right) (2b \cdot 10^n)^{-1}, \text{ the higher powers of } (2b \cdot 10^n)^{-1} \text{ having no effect upon the first digit in the error.}$$

If  $a \geq b$ , there are 2 digits to the left of the decimal point in the quotient and either  $n-2$  or  $n-3$  zeros in the error.

If  $a < b$ , there is 1 digit to the left of the point in the quotient and either  $n-1$  or  $n-2$  zeros in the error.

Only in rare cases will there be as few as  $n-3$  zeros in front of the first digit in the error. To secure this  $\epsilon$  must be greater than  $10^{2-n}$ . This probability will differ for different values of  $n$ . For example, if  $n=4$ , we have as bounding conditions,

$$a > 20b^2 - 0.101b$$

$$b < 1.0 \text{ and } 0.1 < a < 1.0.$$

The ratio of the area bounded by  $a = 20b^2 - 0.101b$ ,  $b = .1$ , and  $a = 1.0$  to the area of the square bounded by  $a = .1$ ,  $a = 1.0$ ,  $b = .1$ , and  $b = 1.0$ , is

$$P = \frac{1}{40} \int_{b=.1}^{b=1.0} (20b^2 - 0.101b) db = 0.0084$$

which is the probability that there would be only  $n - 3$  zeros following the decimal point in the error.

$$(b) \text{ Let } m-n > 1. \text{ Then } \epsilon = \frac{b + a \cdot 10^{m-n}}{b(2b \cdot 10^{n-1})}$$

This situation should be studied for specific values of  $n$ ..

However an approximation may be obtained by letting

$$\epsilon' = \frac{b + a \cdot 10^{m-n}}{2b^2 \cdot 10^n}$$

since subsequent terms in the expansion do not affect the first digit.

If  $a \geq b$ , then  $\epsilon' < \frac{1+10^{m-n+1}}{2b \cdot 10^n} = \frac{10^{m-2n+1}}{2b} + \text{other terms which do not affect the first digit.}$

$$\text{Also } \epsilon' > \frac{1+10^{m-n}}{2b \cdot 10^n} > \frac{10^{m-2n}}{2b} > 0.5(10^{m-2n}).$$

In this case the quotient has  $m-n+1$  digits to the left of the decimal point, while the error has either  $m-2n$ ,  $m-2n+1$  or  $m-2n+2$ . Consequently there are either  $n-1$ ,  $n$ , or  $n+1$  digits free from error in the result.

$$\text{If } a < b, \text{ then } \frac{1+10^{m-n-1}}{2b \cdot 10^n} < \epsilon' < \frac{1+10^{m-n}}{2b \cdot 10^n}$$

In this case the quotient has  $m-n$  digits to the left of the decimal point, while the error has either  $m-2n-1$ ,  $m-2n$ , or  $m-2n+1$ . Again there are either  $n-1$ ,  $n$ , or  $n+1$  digits free from error in the result.

(III) Let  $m < n$ .

Suppose  $m+1 = n$ .

If  $a \geq b$ , the first digit of the quotient is immediately to the right of the decimal point, while there are from  $m-1$  to  $m+1$  zeros between the point and the first digit in the error.

If  $a < b$ , there is one zero between the decimal point and the first digit of the quotient, while in the error there are either  $m$  or  $m+1$  zeros.

*In general, therefore, if there are  $n$  digits in the less reliable of two approximations, there will be either  $n$ ,  $n-1$ , or  $n+1$  digits*

free from error in their quotient. In a few rare cases, a fortuitous combination of digits, discussed later, may throw the error back into the  $n-2$  place. In general the quotient should be rounded off to contain only as many places as there are in the less accurate of the two numbers.

### 5. Square Root.

When a number is in excess by 0.5, the error in its square root is  $(A + 1/2)^{1/2} - A^{1/2}$

$$= 1/4 A^{-1/2} - 1/32 A^{-3/2} + 1/128 A^{-5/2} - \dots + (-1)^{K-1} \frac{1 \cdot 3 \cdot 5 \dots (2K-3)}{2^{2K} \cdot K!} A^{-\frac{2K-1}{2}} + \dots$$

When a number is in defect by 0.5, the error in its square root is  $(A - 1/2)^{1/2} - A^{1/2}$

$$= 1/4 A^{-1/2} - 1/32 A^{-3/2} - 1/128 A^{-5/2} - \dots - \frac{1 \cdot 3 \cdot 5 \dots (2K-3)}{2^{2K} \cdot K!} A^{-\frac{2K-1}{2}} - \dots$$

Obviously the greater error occurs when the number is in defect by 0.5, but in either case we may neglect all terms after the first. Each term can readily be shown to be larger than the term following it, and the ratio of the first term to the second is so large that the second term cannot affect the first digit in the error.

We must consider in turn the case in which  $m$  is even and that in which  $m$  is odd.

(I) Let  $m = 2n$ .

Then  $A = a \cdot 10^{2n}$  and has  $2n$  digits to the left of the decimal point.  $A^{1/2} = a^{1/2} \cdot 10^n$  and has  $n$  digits to the left of the decimal.

Now  $|e'| = 1/4 a^{-1/2} \cdot 10^{-n}$ . But  $1/4 < \frac{1}{4a^{1/2}} \leq \frac{1}{4\sqrt{0.1}} < .791$ .

Since  $0.25 (10^{-n}) < |e'| \leq .791 (10^{-n})$ , the error has  $n$  zeros between its first digit and the decimal point.

Therefore when  $m$  is even, the root contains as many significant figures as the number.

(II) Let  $m = 2n - 1$ .

Then  $A = a \cdot 10^{2n-1} = 10a \cdot 10^{2n-2}$

Then  $A^{1/2} = (10a)^{1/2} \cdot 10^{n-1}$  and has  $n$  digits to the left of the decimal.

$$|\epsilon'| = \frac{1}{4} A^{-1/2} = \frac{10^{1-2n}}{4(10a)^{1/2}}$$

$$0.079 < \frac{1}{4\sqrt{10}} < \frac{1}{4(10a)^{1/2}} < \frac{1}{4} = 0.25$$

$$(.079) 10^{1-2n} < |\epsilon'| < 0.25 (10^{1-2n})$$

The error then will affect the  $n^{\text{th}}$  place to the right of the decimal point, and the number of digits free from error will be  $n+n-1=2n-1$  which was the number of places in the square.

(III) . There is also the case where the decimal point is so placed that the second digit in the last period is not known, as in  $\sqrt{32.4}$  or  $\sqrt{0.46825}$ .

Here  $|\epsilon'| = \frac{5}{2} A^{-1/2}$  . In this case also the number of digits free from error in the root is the number of digits in the original number.

*In general, then, the number of digits free from error in the square root of a number is the number of digits in the number.*

#### 6. Effect of the Error.

The following table will illustrate how an error of  $n$  places may affect either  $n$  or  $n+1$  places in the computation :

Result obtained by computation.....	ERROR IN EXCESS			ERROR IN DEFECT		
	6247	5986	7253	6247	5986	7253
Error .....	33	53	12	33	53	12
True value .....	6214	5933	7241	6280	6039	7265
Computed value, rounded .....	6200	6000	7300	6200	6000	7300
True value, rounded .....	6200	5900	7200	6300	6000	7300

We will now show that the chances are approximately 3 out of 4 that an error of  $n$  digits affects  $n$  and not  $n+1$  places in the result. For convenience we may place the decimal point to the left of the first digit in the error, the position of the decimal point being entirely independent of the number of significant figures in the computation.

Let  $\epsilon$  = error.

$d$  = portion of the number to the right of the decimal point.

$c$  = portion of the number to the left of the decimal point.

$A$  = the true value of the number.

Then  $c + d$  = result of computation.

$A = c + d - \epsilon$  = true value.

We will consider  $\epsilon$  to be positive when the observed value is in excess and negative when it is in defect.

We will consider separately the case where the computed value is in excess and the case where it is in defect.

Suppose the result of computation to be in excess

1. (a) Then if  $d > .5$  and  $\epsilon > d - .5$  } the error will affect  $n+1$   
 (b) or  $d < .5$  and  $\epsilon > d + .5$  } places in the result.
2. (a) If  $d > .5$  and  $\epsilon < d - .5$  } the error will affect only  $n$   
 (b) or  $d < .5$  and  $\epsilon < d + .5$  } places in the result.
3. (a) If  $d > .5$  and  $\epsilon = d - .5$  } the error will affect either  
 (b) or  $d < .5$  and  $\epsilon = d + .5$  }  $n$  or  $n+1$  places depend-  
 (c) or  $d = .5$  and  $\epsilon > 0$  } ing on whether the last  
 digit of  $c$  is odd or even.  
 This is on the assumption  
 of the usual rule, that in  
 rounding off the digit 5  
 the previous digit is made  
 even.

Since the number of digits in  $\epsilon$  is finite, the values of  $d$  and of  $\epsilon$  form discrete series, so that we shall have to think of  $d = .5$  not as an infinitesimal but as a finite portion of the scale, ranging from  $d = .495$  to  $d = .505$  when  $n=2$ , from  $d = .4995$  to  $d = .5005$  when  $n=3$ , etc. If we map the region bounded by  $d=0$ ,  $d=1$ ,  $\epsilon=0$ ,  $\epsilon=1$  the proportions of area representing conditions (1), (2) and (3) represent the respective probabilities of these three sets of

conditions. As  $n$  increases, the width of the strip  $d = .5$  becomes smaller, the probability of (3) becomes smaller, and the probability of (2) approaches  $3/4$ .

When  $n=2$  these areas are respectively

1 (a) and 1 (b) .....	.245025
2 (a) and 2 (b) .....	.735075
3 (a), 3 (b), and 3 (c) .....	.0199
	<hr/> 1.000000

We may assume that the last digit in  $c$  is as likely to be even as to be odd, we may say that the probability that the error will affect  $n+1$  places in the result is slightly more than  $1/4$  when there are two digits in  $e$ . This ratio will approach  $1/4$  if the number of digits in  $e$  increases.

A similar argument holds when the result of computation is in defect.

## 7. Summary of Rules.

On the assumption that an error of  $n$  places affects only  $n$  places in the result we have the following rules:

If the less accurate of two approximate numbers contains  $n$  significant digits, their product and their quotient each contain  $n$  or  $n-1$  significant digits.

The square root of a number contains as many significant figures as the number.

About once in four times, the error will affect one more place than these rules state, for the reasons given in section 6.

# COMBINING TWO PROBABILITY FUNCTIONS

By

WILLIAM DOWELL BATEN,  
*University of Michigan.*

The object of this paper is to show results which arise from combining two probability functions in finding the probability function for the sum of two independent variables. The first part presents the sum function when the probability law for each individual variable is "one-half" of the Pearson Type X law. From this law arise certain ideas concerning the Beta function which are not presented by texts treating this subject.

The second part presents some peculiar probability functions when special laws for the individual variables are considered. Here certain laws with infinite discontinuities are combined.

I. The probability function for the sum of  $n$  variables when each is subject to the function  $e^{-x}$ .

Let the probability that the chance variable  $x_1$  lies in the interval  $(x_1, x_1 + dx_1)$  be to within infinitesimals of higher order  $f(x_1) dx_1$ , and the probability that the chance variable  $x_2$  lies in the interval  $(x_2, x_2 + dx_2)$  be to within infinitesimals of higher order  $g(x_2) dx_2$ , where  $x_1$  and  $x_2$  may have respectively any real value.

By a well known theorem, the probability that the sum,  $x_1 + x_2 = z$ , lies in the interval  $(z, z + dz)$  is, to within infinitesimals of higher order,

$$F(z) dz = \int_{-\infty}^{\infty} f(x_1) \cdot g(z - x_1) dx_1 \cdot dz.$$

Let

$$\begin{aligned} f(x_1) &= e^{-x_1} && \text{for } (0, \infty) \\ &= 0 && \text{elsewhere,} \end{aligned}$$

and

$$\begin{aligned} g(x_2) &= e^{-x_2} && \text{for } (0, \infty) \\ &= 0 && \text{elsewhere.} \end{aligned}$$



According to the above theorem, the probability function for the sum,  $x_1 + x_2 = z$ , is

$$\begin{aligned} F_2(z) &= \int_0^z e^{-x_1} e^{-(z-x_1)} dx_1 \\ &= ze^{-z} \quad \text{for } (0, \infty) \\ &= 0 \quad \text{elsewhere,} \end{aligned}$$

which is a Pearson Type III function. The probability functions or laws for  $x_1$  and  $x_2$  are discontinuous at the origin, while the law for the sum is continuous from minus infinity to plus infinity.

By using  $F_2(y)$  and  $g(x_3)$ , the frequency function for the sum,  $x_1 + x_2 + x_3 = z$ , is

$$\begin{aligned} F_3(z) &= \int_0^z x e^{-x} e^{-(z-x)} dx \\ &= \frac{z^2}{2} e^{-z} \quad \text{for } (0, \infty) \\ &= 0 \quad \text{elsewhere; where } x_1 + x_2 = y. \end{aligned}$$

In general, if the probability function for the individual variable  $x_i$  is

$$\begin{aligned} f_i(x_i) &= e^{-x_i}, \quad \text{for } (0, \infty) \\ &= 0 \quad \text{elsewhere,} \end{aligned}$$

then the probability function for the sum,  $\sum_{i=1}^n x_i = z$  is

$$\begin{aligned} F_n(z) &= (z^{n-1} e^{-z}) / (n-1)! \quad \text{for } (0, \infty) \\ &= 0 \quad \text{elsewhere.} \end{aligned}$$

This is also a Type III law. Others have studied this law and have obtained functions for the sum and the average.<sup>1</sup>

<sup>1</sup> Mayr—Wahrscheinlichkeitsfunktionen und ihre Anwendungen—Monatshefte für Math. und Phys., Vol. 30, 1920. p. 20.

Church—On the mean and squared standard deviation of small samples from any population—Biometrika, Vol. 18, 1926. pp. 321-394.

Irwin—On the frequency distribution of the means of samples from a population having any law of frequency with finite moments, with special reference to Pearson Type II—Biometrika, Vol. 19, 1927. pp. 225-239.

C. C. Craig—Sampling when the parent population is a Pearson Type III—Biometrika, Vol. 21, 1929. pp. 287-293.

A. T. Craig—On the distribution of certain Statistics—Am. Jour. of Math., Vol. 54, No. 2, 1932. pp. 353-366.

Baten—Frequency laws for the sum of  $n$  variables which . . .

The purpose of developing this law for the sum of  $n$  independent variables is to show how certain finite summations are evaluated. An interesting summation arises when  $f$  and  $g$  are interchanged in certain cases. For example the law for the sum,

$\bar{z} = \sum_{i=1}^n x_i$   
is  $F_n(\bar{z})$ , and the law for the sum,

$$\begin{aligned} \bar{z} = \sum_{i=1}^{n+1} x_i \text{ is } \int_0^{\bar{z}} F_n(x) f_1(\bar{z}-x) dx &= \int_0^{\bar{z}} f_1(x) F_n(\bar{z}-x) dx = \\ (a) \quad F_{n+1}(\bar{z}) &= \frac{1}{(n-1)!} \int_0^{\bar{z}} e^{-x} e^{-\bar{z}+x} \left[ \bar{z}^{n-1} + (-1)^1 C_{n-1}^1 \bar{z}^{n-2} x + \dots + (-1)^{n-1} x^{n-1} \right] dx \\ &= e^{-\bar{z}} \bar{z}^n \left[ \sum_{r=0}^{n-1} \frac{(-1)^r C_{n-1}^r}{n+1} \right] / (n-1)! \\ &= 0, \quad \text{elsewhere.} \end{aligned}$$

Since the probability function for the sum of the first  $n+1$  variables, when each is subject to  $f_1$ , is

$$(b) \quad \bar{z}^n e^{-\bar{z}} / n! \quad , \text{ for the positive axis}$$

then (a) and (b) are equal and the summation in the above expression for (a) is equal to  $1/n$ ; hence

$$\sum_{r=0}^{n-1} \frac{(-1)^r C_{n-1}^r}{n+1} = 1/n .$$

If the probability function for the sum of the first  $2n$  variables is obtained by "combining" the probability function for the sum of the first  $n$  variables with the probability function for the sum of the following  $n$  variables, another interesting summation arises. This summation is a Beta function in disguise. For example the probability function for the sum,  $x_1 + x_2 + x_3 + x_4 = \bar{z}$ , is  $\bar{z}^3 e^{-\bar{z}} / 3!$  for positive  $\bar{z}$  and zero elsewhere, and the probability function for the sum,  $x_5 + x_6 + x_7 + x_8 = v$ , is  $v^3 e^{-v} / 3!$  for positive  $v$  and zero elsewhere. The probability function for the sum,  $\bar{z} + v = \sum_{i=1}^8 x_i = w$ ,

$$\begin{aligned} \text{is } F(w) &= \frac{1}{3!3!} \int_0^w e^{-\bar{z}} e^{-w+\bar{z}} \bar{z}^3 (w-3w^2\bar{z} + 3w\bar{z}^2 - \bar{z}^3) d\bar{z} \\ &= \frac{1}{3!3!} e^{-w} (1/4 - 3/5 + 3/6 - 1/7) w^7 ; \text{ for the positive axis} \end{aligned}$$

and zero elsewhere. The quantity in parentheses has for numerators the coefficients of the binomial  $(a-b)^3$ , while the denominators begin with a number greater by one than the exponent of the binomial and increase by unity from term to term. The above form suggests the following integral

$$\sum_{r=0}^3 \frac{(-1)^r C_r}{r+4} = \int_0^1 x^3 (1-x)^3 dx = B(4, 4).$$

In general the probability function for the sum of the first  $2n$  variables, by using the probability function for the sum of the first  $n$  and the probability function for the sum of the following  $n$ , is

$$\frac{w^{2n-1} e^{-w}}{(n-1)!(n-1)!} \sum_{r=0}^{n-1} \frac{(-1)^r C_r}{r+n}, \text{ for } (0, \infty) \text{ and zero elsewhere.}$$

The summation can be written as a definite integral

$$\sum_{r=0}^{n-1} \frac{(-1)^r C_r}{r+n} = \int_0^1 x^{n-1} (1-x)^{n-1} dx = B(n, n) = \frac{\Gamma(n) \cdot \Gamma(n)}{\Gamma(2n)}.$$

If the probability law for the sum of  $n$  independent variables is obtained by combining the probability law for the sum of the first  $s$  variables with the law for the sum of the following  $n-s$  variables the following summation arises which is also equal to a Beta function. This summation is

$$\sum_{r=0}^{n-s-1} \frac{(-1)^r C_r}{s+r} = \int_0^1 x^{s-1} (1-x)^{n-s-1} dx = B(s, n-s).$$

This idea concerning the Beta function appears to be new.

## II. Combining two probability functions.

Combining here shall mean finding the probability function for the sum of the variables from the probability functions of the individual variables. Many peculiar functions arise when various laws are used for the probability functions of the individual variables. This section presents a few of them.

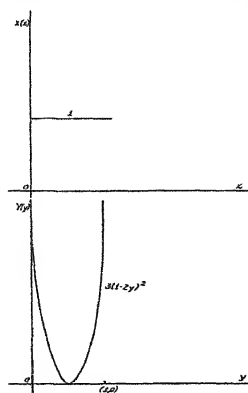
Let

$$f(x) = 1, \text{ for } (0, 1) \text{ and zero elsewhere,}$$

and

$$g(y) = 3(1-2y)^2, \text{ for } (0, 1) \text{ and zero elsewhere.}$$

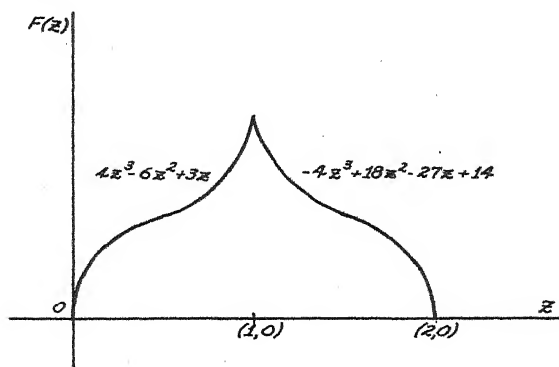
These laws are drawn below. Both have two points of discontinuity.



The probability function for the sum,  $x+y=z$  is

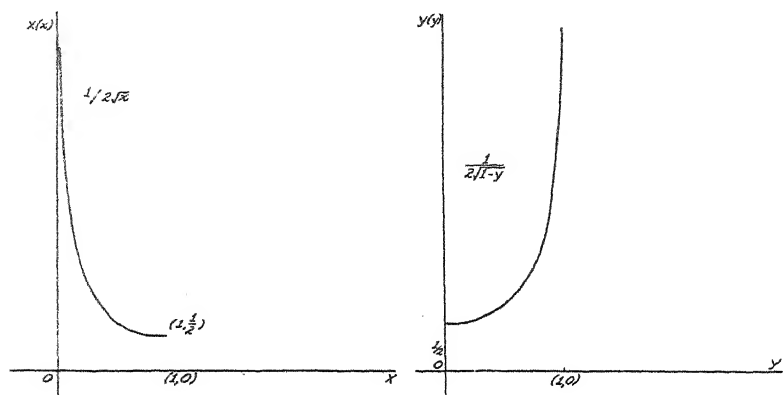
$$F(z) = \begin{cases} (4z^3 - 6z^2 + 3z) & , \text{ for the interval } (0,1) \\ (-4z^3 + 18z^2 - 27z + 14) & , \text{ for } (1,2) \\ 0, & \text{ elsewhere.} \end{cases}$$

$F(z)$  is continuous, symmetrical about the line  $z=1$  with large slope at the points  $(0,0)$ ,  $(1,1)$  and  $(2,0)$ . There is a cusp at  $(1,1)$ .  $F(z)$  is drawn below.



2. Let  $f(x) = \frac{1}{2\sqrt{x}}$ , for  $(0,1)$  and zero elsewhere, and  $g(y) = \frac{1}{2\sqrt{1-y}}$ , for  $(0,1)$  and zero elsewhere. The function  $f(x)$  is the probability function for the square of the variable if the probability law for the variable is unity in  $(0,1)$  and zero elsewhere. The function  $f(x)$  approaches infinity at the origin,

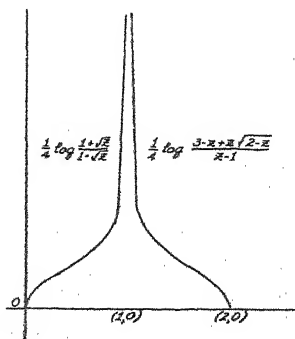
while  $g(y)$  is a similar curve turned in the opposite direction and has an infinite slope at  $(1,0)$ .



The law for the sum,  $x+y = z$  is

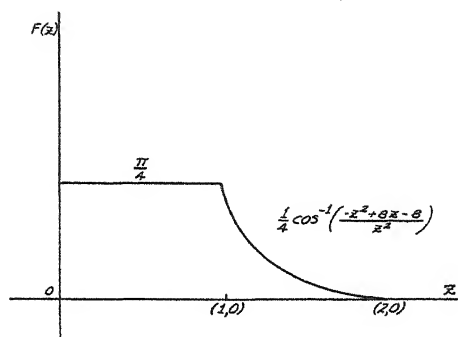
$$F(z) = \begin{cases} \frac{1}{4} \cdot \log \left[ \frac{(1+\sqrt{z})}{(1-\sqrt{z})} \right], & \text{for } (0,1) \\ \frac{1}{4} \cdot \log \left[ \frac{(3-z+\sqrt{2-z})}{(z-1)} \right], & \text{for } (1,2) \\ 0, & \text{elsewhere.} \end{cases}$$

$F(z)$  is somewhat of a surprise for it is equal to zero at the origin and the point  $(2,0)$  and approaches infinity from the right and from the left at the point  $(1,0)$ . The slope of the law for the sum is infinite at the origin and at the points  $(2,0)$  and  $(1,0)$ .  $F(z)$  appears below.



3. Let  $f(x)=1$ , for  $(0,1)$  and zero elsewhere and  $g(y)=1$  for the interval  $(0,1)$  and zero elsewhere; then the probability law of  $w=x^2$  is  $h(w) = \frac{1}{2\sqrt{w}}$ , for the interval  $(0,1)$  and zero elsewhere. Let  $u=y^2$ , then the probability function for  $u$  is  $k(u) = \frac{1}{2\sqrt{u}}$ , for the interval  $(0,1)$  and zero elsewhere. According to the theorem used in part I the probability law for the sum,  $x^2+y^2=z$ , is

$$F(z) = \begin{cases} \pi/4, & \text{for the interval } (0,1) \\ 1/4 \cdot \arccos \frac{-z^2+8z-8}{z^2}, & \text{for } (1,2) \\ 0, & \text{elsewhere. The plot of } F(z) \text{ is below.} \end{cases}$$



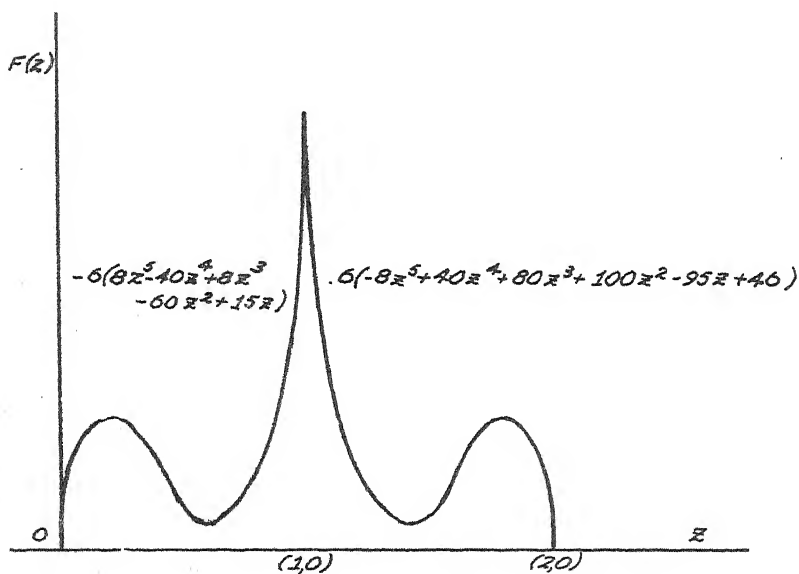
The functions  $h(w)$  and  $k(u)$  are J-shaped functions with infinite slope at the origin and are equal to  $f(x)$  in example 2. The law for the sum of the squares in this case has one point of discontinuity which is at the origin. The function for the sum is constant throughout the interval  $(0,1)$  and is equal to an inverse cosine function throughout the interval  $(1,2)$ .

4. If  $f(x)=3(1-2x)^2$  for the interval  $(0,1)$  and zero elsewhere and  $g(y)=3(1-2y)^2$  for the interval  $(0,1)$  and zero elsewhere, then the law for the sum,  $x+y=z$ , is

$$F(z) = \begin{cases} .6(8z^5-40z^4+80z^3-60z^2+15z), & \text{for the interval } (0,1) \\ .6(-8z^5+40z^4-80z^3+100z^2-95z+46) & \text{for } (1,2) \\ c, & \text{elsewhere.} \end{cases}$$

The function  $F(z)$  has three modes and has its highest point

where one would least expect it, and has large slopes at the origin, and at the points  $(1,0)$ ,  $(2,0)$ . To appreciate the nature of  $F(z)$  here the graphs of the functions for  $x$  and  $y$  should be examined. They are U-shaped curves which are tangent to the horizontal axis at the middle of the interval  $(0,1)$ . See the second figure in 1.  $F(z)$  is shown in the following figure.



# ON THE SYSTEMATIC FITTING OF STRAIGHT LINE TRENDS BY STENCIL AND CALCULATING MACHINE

By

HERBERT A. TOOP'S,  
*Ohio State University*

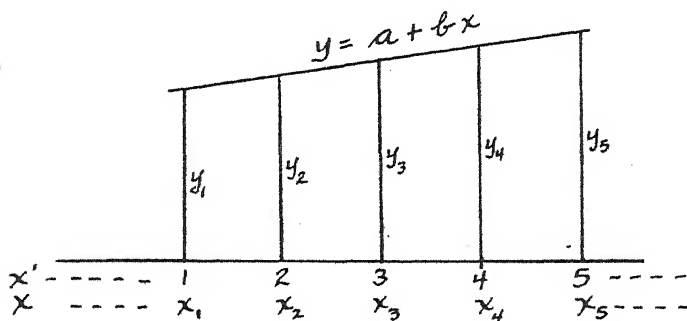
Whenever there is only one plotting point corresponding to  $N$  successive abscissal values equally spaced, it is possible greatly to simplify the fitting of straight lines to the empirical observations. Let the  $N$  several abscissal values (ordinarily time) be

$$x_1, x_2, x_3, \dots, x_N.$$

Let these several  $x$  values be replaced by a series of transmuted steps,  $x'_i$  ( $i=1, 2, 3, \dots, N$ ).

Let the several corresponding ordinates be  $y_1, y_2, y_3, \dots, y_N$ . The situation is represented in Figure 1.

FIG. 1. Illustrating the Notation Employed.



Letting the equation of the fitted straight line be

$$(1) \quad y = a + bx'.$$

it is well known that the solutions, by least squares, for the two constants are,

$$(2) \quad a = \frac{\sum y \cdot \sum (x')^2 - \sum x' \sum x'y}{N \sum (x')^2 - (\sum x')^2},$$

and

$$(3) \quad b = \frac{N \sum x'y - \sum x' \sum y}{N \sum (x')^2 - (\sum x')^2}.$$



Also that, inasmuch as the  $x'$  coordinates are an arithmetical series, we may substitute in the above for  $\sum x'$  and  $\sum (x')^2$  as follows:

$$(4) \quad \sum x' = \frac{1}{2} N(N+1),$$

$$(5) \quad \sum (x')^2 = \frac{1}{6} N(N+1)(2N+1),$$

thus yielding,

$$(6) \quad a = \frac{(4N+2)\sum y - 6\sum x'y}{N(N-1)},$$

$$(7) \quad b = \frac{12\sum x'y - 6(N+1)\sum y}{N(N^2-1)}.$$

which equations, if of infrequent usage, are highly serviceable. It is possible, however, to proceed to the derivation of formulae still more useful for systematic fitting of straight line trends. Thus, there being only one ordinate to each abscissal value as assumed,

$$(8) \quad \sum y = y_1 + y_2 + y_3 + \cdots + y_N,$$

$$(9) \quad \sum x'y = 1y_1 + 2y_2 + 3y_3 + \cdots + Ny_N.$$

It will be observed further that the denominator  $N(N-1)$  of (6) is invariably an even number, and therefore exactly divisible by 2.<sup>1</sup> Substituting (8) and (9) in (6), and then multiplying both numerator and denominator by  $\frac{1}{2}$ , we obtain

$$(10) \quad a = \frac{(2N+1)(y_1 + y_2 + \cdots + y_N) - 3(1y_1 + 2y_2 + \cdots + Ny_N)}{N(N-1)},$$

an equation which is a function only of the several ordinates and of  $N$ . Furthermore, this equation when solved for specific values of  $N$  leads to a system of equations remarkably simple; and moreover one easily extended indefinitely.

Thus, when

$$(11) \quad N=2, \quad a_2 = \frac{1}{1} (2y_1 - y_2)$$

<sup>1</sup> The desiderata are: 1. To obtain a formula which shall obtain as small multipliers of the several  $Y$ 's as possible, consistent with

2. Integral multipliers, and

3. An integral numerator and denominator for the  $A$  and  $B$  coefficients.

$$(12) \quad \text{When } N = 3, \quad a_3 = \frac{1}{3}(4y_1 + 1y_2 - 2y_3)$$

$$(13) \quad \text{When } N = 4, \quad a_4 = \frac{1}{6}(6y_1 + 3y_2 + 0y_3 - 3y_4)$$

etc., etc.

The symmetry of arrangement is more readily grasped if the several coefficients of the  $y'_s$ , and the denominators,  $D_a$ , be collected into an orderly table thus (Figure 2) :—

FIG. 2. Systematic Solution of Equation (10), for Specific Values of , for Finding  $a$  .

$$y = a + bx'$$

RULE: Extend and cumulate the successive  $y$ 's by the stencil multipliers of the row of the table appropriate to the problem (determined by  $N$  ) in question. Divide the accumulated sum by the denominator,  $D_a$ , of the same row. The resulting quotient is  $a$  .

N	$D_a$	Multiplier of the ordinate:—											
		$y_1$	$y_2$	$y_3$	$y_4$	$y_5$	$y_6$	$y_7$	$y_8$	$y_9$	$y_{10}$	$y_{11}$	$y_{12}$
2	1	2	-1										
3	3	4	1	-2									
4	6	6	3	0	-3								
5	10	8	5	2	-1	-4							
6	15	10	7	4	1	-2	-5						
7	21	12	9	6	3	0	-3	-6					
8	28	14	11	8	5	2	-1	-4	-7				
9	36	16	13	10	7	4	1	-2	-5	-8			
10	45	18	15	12	9	6	3	0	-3	-6	-9		
11	55	20	17	14	11	8	5	2	-1	-4	-7	-10	
12	66	22	19	16	13	10	7	4	1	-2	-5	-8	-11

Having such a table at hand it is obvious that  $a$  may be determined quickly by

1. Simply choosing the appropriate row of multipliers for the number,  $N$ , of successive plotting points available<sup>2</sup>; and

2. Extending the several  $y'_s$  by the appropriate multipliers, most conveniently done by calculating machine;

3. Dividing the sum so obtained by the appropriate divisor,  $D_a$  .

<sup>2</sup> Obviously if any plotting point intermediate between  $y_1$  and  $y_N$  is missing it must be supplied (by interpolation) before employing this method.

# SYSTEMATIC FITTING OF STRAIGHT LINE

The multipliers may be extended indefinitely for larger and larger values of  $N$  by simply noting that the diagonal marginal row increases by the successive addition of  $-1$ , while columns increase by the successive addition of  $2$ ; and rows decrease by the successive addition of  $-3$ . The denominators have a constant second order difference,  $\Delta^2 = 1$ , and consequently may be prolonged readily.

Let us now return to  $\ell$ , equation (7). The denominator  $N(N^2-1)$  is always divisible by 6. Hence, substituting (8) and (9) in (7) and dividing both numerator and denominator by 6, we obtain,

$$(14) \ell = \frac{2(y_1 + 2y_2 + 3y_3 + \dots + Ny_N) - (N+1)(y_1 + y_2 + y_3 + \dots + y_N)}{\frac{N(N^2-1)}{6}}$$

In like manner, this equation when solved for specific values of  $N$  leads to a systematic series of equations:

$$(15) \ell_2 = \frac{1}{1} (-y_1 + y_2) \quad (\text{where } N=2)$$

$$(16) \ell_3 = \frac{1}{4} (-2y_1 + 0y_2 + 2y_3) \quad (\text{where } N=3)$$

$$(17) \ell_4 = \frac{1}{10} (-3y_1 - 1y_2 + 1y_3 + 3y_4) \quad (\text{where } N=4).$$

The corresponding table yields Figure 3.

FIG. 3. Systematic Solution of Equation (14), for Specific Values of  $N$ , for Finding  $\ell$ .

$$y = a + \ell x'.$$

RULE: Extend and cumulate the successive  $y^2$ s by the stencil multipliers of the row of the table appropriate to the problem (determined by  $N$ ) in question. Divide the accumulated sum by the denominator,  $D_4$ , of the same row. The resulting quotient is  $\ell$ .

N	$D_4$	Multiplier of the ordinate:—											
		$y_1$	$y_2$	$y_3$	$y_4$	$y_5$	$y_6$	$y_7$	$y_8$	$y_9$	$y_{10}$	$y_{11}$	$y_{12}$
2	1	-1	1										
3	4	-2	0	2									
4	10	-3	-1	1	3								
5	20	-4	-2	0	2	4							
6	35	-5	-3	-1	1	3	5						
7	56	-6	-4	-2	0	2	4	6					
8	84	-7	-5	-3	-1	1	3	5	7				
9	120	-8	-6	-4	-2	0	2	4	6	8			
10	165	-9	-7	-5	-3	-1	1	3	5	7	9		
11	220	-10	-8	-6	-4	-2	0	2	4	6	8	10	
12	286	-11	-9	-7	-5	-3	-1	1	3	5	7	9	11

The extension of this table is readily made by observing that the diagonal marginal row increases by adding 1; the columns, by adding -1, and the rows, by adding 2; while the denominator has a constant third order difference,  $\Delta^3 = 1$ .

For hand computations these two tables, Figures 2 and 3, are undoubtedly simplest because the multipliers are smallest. If, however, a calculating machine is available, the magnitude of the multipliers is of relatively small moment if anything is to be gained by using different multipliers. It is obvious, for example, that the several multipliers of a row may be divided by the appropriate denominator, the resulting decimal multipliers, to replace the present integral multipliers, being presented in tables of  $N$  columns or sections.

An even more useful set of tables for general purposes may be derived by reducing equations (10) and (14) to a common (integral) denominator, so that the same denominator may be employed for calculating both  $\alpha$  and  $\ell$ .

We may obtain the least common denominator,  $\frac{N(N^2-1)}{2}$  by multiplying equation (10) by  $\frac{N+1}{N+1}$ ; and, equation (14), by multiplying by  $\frac{3}{3}$ , thus:

$$(18) \quad \alpha = \frac{(N+1)[(2N+1)(y_1+y_2+\dots+y_N)-3(y_1+2y_2+\dots+Ny_N)]}{\frac{N(N^2-1)}{2}}$$

$$(19) \quad \ell = \frac{6(y_1+2y_2+\dots+Ny_N)-3(N+1)(y_1+y_2+\dots+y_N)}{\frac{N(N^2-1)}{2}}$$

Accordingly, it follows that if the three following changes be effected, we shall have an integral system:

1. The previous table values of  $\alpha$  to be multiplied by  $(N+1)$  of the row in question throughout.

2. The previous table values of  $\ell$  to be multiplied by 3 throughout.

3. The common denominator,  $\frac{N(N^2-1)}{2}$ , of any row in question to be made to be 3 times the previous denominator of  $\ell$  of the row.

The two sets of multipliers may now be combined into one systematic stencil (Figure 4) with a common denominator  $D$  or common reciprocal,  $\frac{1}{D}$ . The directions for using this stencil are as follows:—

1. Count the number of plotting points.
2. Find the row of the stencil having the same number of plotting points, ( $N$ ).
3. Record the  $y$  values for the successive plotting points in the little rectangles of the row just located.
4. Using a calculating machine, obtain the summation of the extensions of the several  $y$ -values by the multipliers just *immediately above*, employing a fixed decimal point.
5. Divide the sum just found by the divisor,  $D$ , at the left hand of the row. The result is  $\ell$ .
6. Similarly obtain the summation of the extensions of the  $y$ 's by the multipliers of the several respective windows *immediately beneath*, again employing a fixed decimal point.
7. Divide the sum thus obtained by the same divisor,  $D$ . The result is  $a$ .
8. Substitute values of  $a$  and  $\ell$  in

$$(1) \quad y = a + \ell x'.$$

9. If we summate (1) we obtain the checking equation,

$$(20) \quad \sum y = \frac{1}{2} [2Na + N(N+1)\ell],$$

since  $\sum x' = \frac{1}{2} N(N+1)$ .

Now, let us employ the revised stencil on a problem (of perfect fit):—

$x$ (Age)	$y$ (Attainment)	$x'$
3.5	12.72	1
5.5	22.45	2
7.5	32.18	3
9.5	41.91	4
11.5	51.64	5
13.5	61.37	6
15.5	71.10	7 = N
$\Sigma x = 66.5$	$293.37 = \Sigma y$	

FIG. 4. Revised Stencil for Solving Formulae (19) and (18) for  $b$  and  $a$ , respectively,  $y = a + bx'$ .

$$(19) \quad b = \frac{6(1y_1 + 2y_2 + \dots + Ny_N) - 3(N+1)(y_1 + y_2 + \dots + y_N)}{\frac{N(N^2-1)}{2}}$$

$$(18) \quad a = \frac{(N+1)[(2N+1)(y_1 + y_2 + \dots + y_N) - 3(1y_1 + 2y_2 + \dots + Ny_N)]}{\frac{N(N^2-1)}{2}}$$

$$D = \frac{N(N^2-1)}{2}$$

Multiplier of Ordinate No.:—

N    D    1    2    3    4    5    6    7    8    9    10    11    12

2

	-3	3
3	<input type="text"/>	<input type="text"/>
	6	-3

3

	-6	0	6
12	<input type="text"/>	<input type="text"/>	<input type="text"/>
	16	4	-8

4

	-9	-3	3	9
30	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
	30	15	0	-15

5

	-12	-6	0	6	12
60	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
	48	30	12	-6	-24

FIG. 4—Continued

6	105	-15	-9	-3	3	9	15												
		70	49	28	7	-14	-35												
7	168	-18	-12	-6	0	6	12	18											
		96	72	48	24	0	-24	-48											
8	252	-21	-15	-9	-3	3	9	15	21										
		126	99	72	45	18	-9	-36	-63										
9	360	-24	-18	-12	-6	0	6	12	18	24									
		160	130	100	70	40	10	-20	-50	-80									
10	495	-27	-21	-15	-9	-3	3	9	15	21	27								
		198	165	132	99	66	33	0	-33	-66	-99								
11	660	-30	-24	-18	-12	-6	0	6	12	18	24	30							
		240	204	168	132	96	60	24	-12	-48	-84	-120							
12	858	-33	-27	-21	-15	-9	-3	3	9	15	21	27	33						
		286	247	208	169	130	91	52	13	-26	-65	-104	-143						

Since the X-coordinates are replaced, for computation, by the series, the following transmuting equation prevails:

$$(21) \quad x' = .5x - .75.$$

The stencil set up, employed for seven  $y'_s$ , is:—

	-18	-12	-6	0	6	12	18
D=168	12.72	22.45	32.18	41.91	51.64	61.37	71.10
	96	72	48	24	0	-24	-48

$$b = \frac{1}{168} [-18(12.72) - 12(22.45) - \dots + 18(71.10)] = 9.730$$

$$a = \frac{1}{168} [96(12.72) + 72(22.45) + \dots - 48(71.10)] = 2.990$$

whence:

$$y = 2.990 + 9.730 x'$$

At this stage, application of formula (20) proves the correctness of this equation.

Now substitute for  $x'$  its equivalent  $(.5x+.75)$  and

$$y = 4.865x - 4.3075,$$

which may be checked by summing,

$$\sum y = 4.865 \sum x - 4.3075 N = 4.865(66.5) - 4.3075(7);$$

i.e.  $293.37 = 293.37$  . The check holds.



# STATISTICAL ANALYSIS OF ONE-DIMENSIONAL DISTRIBUTIONS

By

ROBERT SCHMIDT

The present research is to be considered as a contribution to a range of science in which the pioneer work has been done by K. PEARSON. The method for analysing statistical distributions to be developed here differs in principle—as far as the author can see—from the known ones. The mathematical resources are all well known and so simple that their deduction *ab ovo* could be carried through on a few pages; hence this investigation is intelligible to anyone who remembers his mathematical knowledge acquired at school.

The main resource consists of the process of orthogonalization, fundamental in the theory of integral equations. The central idea characterizing the following is, not to deal with a frequency function itself, nor with its integral function, but with the *inverse* of the integral function. The general scope will be given in No. 3.

The author is indebted to his wife and to Mr. J. L. K. GIFFORD, M.A., of Queensland University for kind help in revising the English text.

## 1. DESIGNATIONS AND GENERAL ASSUMPTIONS

A curve  $y = \varphi(x)$ ,  $(-\infty < x < +\infty)$  shall be called a “frequency curve”, the function  $\varphi(x)$  a “frequency function”, if  $\varphi(x)$  satisfies the following conditions:

1.  $\varphi(x) \geq 0$   $(-\infty < x < +\infty)$
2. The moments  $\mu_k = \int_{-\infty}^{+\infty} x^k \varphi(x) dx$  exist for  $k = 0, 1, \dots$  <sup>1</sup>
3.  $\mu_0 = 1$ .

For our purposes it is convenient—though not necessary—to

<sup>1</sup> In this paper we shall not have to make use of the second condition (except in the special case  $k=0$ ); in further notes, too, the condition will never be applied to its full extent.

add a fourth condition which it is simplest to formulate by using the function

$$\phi(x) = \int_{-\infty}^x \varphi(t) dt.$$

This function is constantly increasing in  $-\infty < x < +\infty$ , and we have

$$\lim_{x \rightarrow -\infty} \phi(x) = 0, \quad \lim_{x \rightarrow +\infty} \phi(x) = 1$$

The fourth condition is to guarantee that  $\phi(x)$  assumes every value from  $0 < y < 1$  just once, so that  $\phi(x)$  possesses a unique inverse function in the ordinary sense.

4. a)  $\phi(x)$  is continuous

b) At every  $x$  where  $0 < \phi(x) < 1$ ,  $\phi(x)$  is increasing (strictly speaking), that is: From  $x' < x < x''$  it always follows that  $\phi(x') < \phi(x) < \phi(x'')$ .

When the conditions 1 - 4 are fulfilled, let us denote  $\phi(x)$  as the "frequency integral" of the frequency function  $\varphi(x)$ .

Then there exists one and only one function  $\psi(y)$  in  $0 < y < 1$ , satisfying  $\psi[\phi(x)] \equiv x$  ( $0 < \phi(x) < 1$ ) and  $\psi(y)$  is called the *inverse function* of  $\phi(x)$ . This function  $\psi(y)$  is continuous and constantly increasing (strictly speaking), and therefore possesses a unique inverse, namely  $\phi(x)$ :

$$\phi[\psi(y)] \equiv y \quad (0 < y < 1).$$

We give here some special examples of frequency curves.

I. The "Step Curve".

$$y = \varphi(x) = \begin{cases} 1 & \text{in } 0 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

The moments are  $\mu_k = \frac{1}{k+1}$  ( $k = 0, 1, 2, \dots$ ).

The frequency integral is

$$\phi(x) = \begin{cases} 0 & \text{in } -\infty < x < 0 \\ x & \text{in } 0 \leq x < 1 \\ 1 & \text{in } 1 \leq x < +\infty \end{cases}$$

The inverse is

$$\psi(y) = y \quad (0 < y < 1)$$

## II. The Normal Law of Error.

$$y = \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},$$

with the moments

$$\mu_k = \begin{cases} \frac{(2n)!}{2^n \cdot \pi!} & \text{for } k = 2n. \\ 0 & \text{for } k = 2n+1 \quad (n=0, 1, \dots) \end{cases}$$

and the frequency integral

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt.$$

There are a number of tables of the numerical values of this function. Of course these tables can be used to compute the values of  $\psi(y)$ . Considering the fact that, for our purposes, the values of  $\psi(y)$  will often be required for simple rational arguments only, it seems useful to have tables which are converse to those just quoted, that is to say, the tabulated entry of which is  $x = \psi(y)$  and the argument  $y = \phi(x)$ . Such tables have been calculated by KELLEY and WOOD (Statistical Method, New York 1924; Appendix C).

## III. The Laplace Curve.

$$y = \varphi(x) = \frac{1}{2} e^{-|x|}$$

$$\mu_k = \begin{cases} (2n)! = k! & \text{for } k = 2n \\ 0 & \text{for } k = 2n+1 \quad (n=0, 1, \dots) \end{cases}$$

$$\phi(x) = \begin{cases} \frac{1}{2} e^x & \text{in } -\infty < x < 0 \\ 1 - \frac{1}{2} e^{-x} & \text{in } 0 \leq x < +\infty \end{cases}$$

$$\psi(y) = \begin{cases} \log y + \log 2 & \text{in } 0 < y < \frac{1}{2} \\ -\log(1-y) - \log 2 & \text{in } \frac{1}{2} \leq y < 1 \end{cases}$$

## IV. The "Tine Curve".

$$y = \varphi(x) = \begin{cases} 0 & \text{in } -\infty < x < -1 \\ 1+x & \text{in } -1 \leq x < 0 \\ 1-x & \text{in } 0 \leq x < +1 \\ 0 & \text{in } +1 \leq x < +\infty \end{cases}$$

$$\phi(x) = \begin{cases} 0 & \text{in } -\infty < x < -1 \\ \frac{1}{2}(1+x)^2 & \text{in } -1 \leq x < 0 \\ 1 - \frac{1}{2}(1-x)^2 & \text{in } 0 \leq x < +1 \\ 1 & \text{in } +1 \leq x < +\infty \end{cases}$$

$$\psi(y) = \begin{cases} -1 + \sqrt{2y} & \text{in } 0 < y < 1/2 \\ 1 - \sqrt{2-2y} & \text{in } 1/2 \leq y < 1 \end{cases}$$

## 2. EKKÉ'S "BEST VALUES"

A. EKKÉ, in his Kiel dissertation (to appear), deals with the following question among others: Suppose a frequency function  $\varphi(x)$  and a natural number  $n$  given. Which one among all systems of  $n$  values  $x_1, \dots, x_n$  might be considered the "best"?—To give an answer to this question, EKKÉ divides the total  $x$ -axis into  $n$  parts  $I_1, \dots, I_n$  with the separating points  $\alpha_1, \dots, \alpha_{n-1}$  in such a manner that

$$\int_{-\infty}^{\alpha_1} \varphi(x) dx = \int_{\alpha_1}^{\alpha_2} \varphi(x) dx = \dots = \int_{\alpha_{n-1}}^{+\infty} \varphi(x) dx = \frac{1}{n}.$$

Evidently this is possible in one and only one manner, and we have

$$\alpha_1 = \psi\left(\frac{1}{n}\right), \quad \alpha_2 = \psi\left(\frac{2}{n}\right), \quad \dots, \quad \alpha_{n-1} = \psi\left(\frac{n-1}{n}\right)$$

Each of the parts  $I_1, \dots, I_n$  should contain exactly one value of the system. Furthermore it seems reasonable to fix every point  $x_\nu$  within its interval  $I_\nu$  by the conditions

$$\int_{-\infty}^{x_1} \varphi(x) dx = \int_{x_1}^{x_2} \varphi(x) dx, \quad \dots, \quad \int_{x_{n-1}}^{x_n} \varphi(x) dx = \int_{x_n}^{+\infty} \varphi(x) dx.$$

This also can be done in one and only one way. Let us designate these "best values" by  $\xi_1, \dots, \xi_n$ . We have

$$(1) \quad \xi_1 = \psi\left(\frac{1}{2n}\right), \quad \xi_2 = \psi\left(\frac{2}{2n}\right), \quad \dots, \quad \xi_n = \psi\left(\frac{2n-1}{2n}\right).$$

Concerning the best values, ЕККЕ proves two theorems which accentuate the rationality of the definition. If  $x_1 \leq \dots \leq x_n$  are values arranged according to magnitude, and

$$S(x; x_1, \dots, x_n) = \begin{cases} 0 & \text{in } -\infty < x < x_1, \\ \frac{\nu}{n} & \text{in } x_\nu \leq x < x_{\nu+1} \quad (\nu=1, \dots, n-1) \\ 1 & \text{in } x_n \leq x < +\infty. \end{cases}$$

the following theorem holds:

"There is one and only one system  $x_1, \dots, x_n$  for which

$$\int_{-\infty}^{+\infty} \{ \phi(x) - S(x; x_1, \dots, x_n) \}^2 dx$$

assumes a minimum, and this system is  $x_1 = \xi_1, \dots, x_n = \xi_n$ ."

This theorem also holds if the exponent 2 is replaced by an arbitrary positive number.—Furthermore:

"There is one and only one set  $x_1, \dots, x_n$  for which the lowest upper boundary of

$$| \phi(x) - S(x; x_1, \dots, x_n) |$$

assumes a minimum, and this set again is identical with  $\xi_1, \dots, \xi_n$ ."

For normalizing purposes ЕККЕ considers, together with a given frequency function  $\phi(x)$ , the totality of the frequency functions which result by linear transformations of the argument, i.e. which result by translations and dilatations in the direction of the  $x$ -axis (or by choosing new origins and new units of measurement). With an arbitrary  $\beta$ , and  $\alpha > 0$ , we have to form

$$\tilde{\phi}(x) = \frac{1}{\alpha} \phi\left(\frac{1}{\alpha}(x-\beta)\right),$$

the first factor  $\frac{1}{\alpha}$  being required in order to comply with condition 3. The frequency integral corresponding to  $\tilde{\phi}(x)$  is

$$\tilde{\phi}(x) = \phi\left(\frac{1}{\alpha}(x-\beta)\right),$$

and the inverse

$$\tilde{\psi}(y) = \alpha \psi(y) + \beta.$$

Due to this simple relation between  $\psi(y)$  and  $\tilde{\psi}(y)$ , we have evidently, if  $\xi_1, \dots, \xi_n$  designate the best values of  $\tilde{\phi}(x)$ ,

$$\tilde{\xi}_1 = \alpha \xi_1 + \beta, \quad \tilde{\xi}_2 = \alpha \xi_2 + \beta, \quad \dots, \quad \tilde{\xi}_n = \alpha \xi_n + \beta.$$

This fact can be used to pick out from the multitude of functions  $\tilde{\varphi}(x)$  a distinct specimen, and then to operate with its best values only. It is easy to show in a direct manner that there is exactly one specimen in the multitude which complies with the additional conditions  $\mu_1 = 0, \mu_2 = 1$ .

### 3. THE STARTING POINT. GENERAL SCOPE.

But the proof of the fact just mentioned can be given indirectly too by considering the inverses  $\tilde{\psi}(y)$ , and it is this way which gives the starting point of our further developments. Indeed, if we introduce — for simplicity — Stieltjes integrals, the conditions  $\mu_1 = 0, \mu_2 = 1$  mean

$$\int_{-\infty}^{+\infty} x \, d\tilde{\varphi}(x) = 0, \quad \int_{-\infty}^{+\infty} x^2 \, d\tilde{\varphi}(x) = 1,$$

and by the substitution  $x = \tilde{\psi}(y)$  we get

$$\begin{aligned} \text{or} \quad & \int_0^1 \tilde{\psi}(y) \, dy = 0, \quad \int_0^1 \tilde{\psi}^2(y) \, dy = 1 \\ & \int_0^1 (\beta + \alpha \psi(y)) \, dy = 0, \quad \int_0^1 (\beta + \alpha \psi(y))^2 \, dy = 1. \end{aligned}$$

Let us put

$$\text{and} \quad \psi_0(y) = 1, \quad \psi_1(y) = \psi(y)$$

$$X_0(y) = \alpha_0 \psi_0(y)$$

$$X_1(y) = \beta_1 \psi_0(y) + \tau_1 \psi_1(y).$$

Then our conditions are equivalent to the following demand: Find coefficients  $\alpha_0; \beta_1, \tau_1$  ( $\alpha_0 > 0, \tau_1 > 0$ ) in such a manner that

$$\int_0^1 X_0^2(y) \, dy = 1, \quad \int_0^1 X_0(y) X_1(y) \, dy = 0, \quad \int_0^1 X_1^2(y) \, dy = 1.$$

We add: The functions  $\psi_0(y)$  and  $\psi_1(y)$  are linearly independent, i.e.  $\lambda_0 \psi_0(y) + \lambda_1 \psi_1(y) \equiv 0$  cannot hold except for  $\lambda_0 = \lambda_1 = 0$ .

Now it is obvious that our demand represents a special case of the general problem as follows: *Given a set of linearly independent continuous functions  $\psi_0(y), \psi_1(y), \dots, \psi_k(y)$  ( $0 < y < 1$ ). The scheme*

of coefficients

$$\begin{array}{ccccccc} \beta_{00} & 0 & 0 & \dots\dots\dots 0 \\ \beta_{10} & \beta_{11} & 0 & \dots\dots\dots 0 \\ \hline \beta_{k0} & \beta_{k1} & \beta_{k2} & \dots\dots\dots \beta_{kK} \end{array}$$

satisfying the additional conditions  $\beta_{00} > 0, \beta_{11} > 0, \dots, \beta_{KK} > 0$ , shall be chosen so that the functions

$$Z_0(y) = \beta_{00} \psi_0(y)$$

$$Z_1(y) = \beta_{10} \psi_0(y) + \beta_{11} \psi_1(y)$$

-----

$$Z_k(y) = \beta_{k0} \psi_0(y) + \beta_{k1} \psi_1(y) + \dots + \beta_{kK} \psi_K(y)$$

form a normalized orthogonal system, i.e.

$$\int_0^1 Z_p(y) \cdot Z_q(y) dy = \begin{cases} 1 & \text{for } p = q \\ 0 & \text{for } p \neq q \end{cases}$$

It is well known that there is one and only one suitable scheme, and it is furnished by the so-called *process of orthogonalization*. Furthermore it is well known that the process of orthogonalization is intimately connected with another problem: *An arbitrary continuous function  $F(y)$  given, to determine the coefficients  $c_0, \dots, c_K$*

so that  $\int_0^1 \{F(y) - [c_0 \psi_0(y) + \dots + c_K \psi_K(y)]\}^2 dy$

assumes a minimum.

Concerning frequency functions, we are led — by pursuing this line—to a general theory of curve types; an account of the results to be obtained will be given in a future article.

Concerning our analysis of statistical data, we do not intend to use from a given frequency function more than its best values. More precisely: *we intend to replace the frequency function by its*

best values. Our *modus procedendi* now results by analogy: we have to deal with systems of values (vectors)

$$\begin{array}{c} (u_{11}, u_{12}, \dots, u_{1n}) \\ (u_{21}, u_{22}, \dots, u_{2n}) \\ \text{-----} \\ (u_{k1}, u_{k2}, \dots, u_{kn}) \end{array}$$

which are linearly independent (see No. 5). We have to employ the process of orthogonalization, which gives a normalized orthogonal system (see No. 6)

$$\begin{array}{c} (w_{11}, w_{12}, \dots, w_{1n}) \\ (w_{21}, w_{22}, \dots, w_{2n}) \\ \text{-----} \\ (w_{k1}, w_{k2}, \dots, w_{kn}) \end{array}$$

and we have to direct our attention to the sums of the form

$$\frac{1}{n} \sum_{\nu=1}^n \{ u_{\nu} - (c_1 u_{1\nu} + \dots + c_k u_{k\nu}) \}^2,$$

or better

$$\frac{1}{n} \sum_{\nu=1}^n \{ u_{\nu} - (l_1 w_{1\nu} + \dots + l_k w_{k\nu}) \}^2.$$

Finally we have to introduce the special set of vectors:

$$\begin{array}{c} (1, 1, \dots, 1) \\ (\xi_1, \xi_2, \dots, \xi_n) \\ \text{-----} \\ (\xi_1^{k-1}, \xi_2^{k-1}, \dots, \xi_n^{k-1}) \end{array}$$

where  $\xi_1, \dots, \xi_k$  designate the best values of a frequency function.

We are now in the position to characterize the direction of our research in general words: *A statistical analysis of distributions as an application of the theory of orthogonal systems, based upon the best values of a given frequency function.*

#### 4. VECTORS

For our purpose it is convenient to make use of the notations



and simplest operations of vector analysis. If  $u_1, u_2, \dots, u_n$  are a set of numbers, we take the symbol  $(u_1, \dots, u_n)$  as an individual, call it a *vector*, and designate it by a gothic letter:

$$\check{M} = (u_1, \dots, u_n).$$

Equality of two vectors  $\check{M} = (u_1, \dots, u_n)$  and  $\check{N} = (v_1, \dots, v_n)$  is defined by

$$u_1 = v_1, \quad u_2 = v_2, \quad \dots, \quad u_n = v_n,$$

and is written  $\check{M} = \check{N}$ . The products of a number  $c$  with a vector  $\check{M}$  are defined by

$$c\check{M} = (cu_1, \dots, cu_n)$$

$$\check{M}c = (u_1c, \dots, u_nc);$$

the sum of two vectors  $\check{M}$  and  $\check{N}$  by

$$\check{M} + \check{N} = (u_1 + v_1, \dots, u_n + v_n).$$

Evidently we have

$$c\check{M} = \check{M}c$$

and

$$(\check{M} + \check{N}) + \check{N} = \check{M} + (\check{N} + \check{N}).$$

Hence we may omit the brackets, and the sum of three or more vectors has a definite sense. More general, the meaning of the expression

$$c_1\check{M}_1 + \dots + c_k\check{M}_k$$

is clear. The product of two vectors is (somewhat differently from the customary way) defined as a NUMBER, viz.

$$\check{M}\check{N} = \frac{1}{n} (u_1v_1 + \dots + u_nv_n),$$

and we have

$$\check{M}\check{N} = \check{N}\check{M}$$

$$(\check{M} + \check{N})\check{N} = \check{M}\check{N} + \check{N}\check{N}.$$

But in general the vectors  $(\check{M}\check{N})\check{N}$  and  $\check{M}(\check{N}\check{N})$  are entirely different.

Let us put

$$\check{O} = (0, 0, \dots, 0).$$

Every vector  $\check{M}$  satisfies  $\check{M}^2 = \check{M}\check{M} \geq 0$ , and  $\check{M} = \check{O}$  is the only vector for which  $\check{M}^2 = 0$  holds. — Whenever the square root of the square of a vector,

$$\sqrt{\check{M}^2} = \sqrt{\frac{u_1^2 + \dots + u_n^2}{n}},$$

is met with, we always mean the positive value.

### 5. LINEAR INDEPENDENCE

A set of vectors  $\check{M}_1, \dots, \check{M}_K$  is said to be *linearly independent* if the equation

$$\lambda_1 \check{M}_1 + \lambda_2 \check{M}_2 + \dots + \lambda_K \check{M}_K = \sigma$$

does not hold except for  $\lambda_1 = \lambda_2 = \dots = \lambda_K = 0$ . Otherwise the vectors are said to be *linearly dependent*. If the vectors

$$\check{M}_1, \check{M}_2, \dots, \check{M}_K$$

are linearly independent, all the more the same is true for every partial system. Especially:

$$\check{M}_1 \neq \sigma, \quad \check{M}_2 \neq \sigma, \quad \dots, \quad \check{M}_K \neq \sigma.$$

THEOREM 1. "Let  $\check{M}_1, \dots, \check{M}_K$  be linearly independent; form the vectors

$$(2) \quad \begin{cases} \check{M}_1^* = a_{11} \check{M}_1 \\ \check{M}_2^* = a_{21} \check{M}_1 + a_{22} \check{M}_2 \\ \text{-----} \\ \check{M}_K^* = a_{K1} \check{M}_1 + a_{K2} \check{M}_2 + \dots + a_{KK} \check{M}_K \end{cases}$$

and suppose

$$a_{11} \neq 0, \quad a_{22} \neq 0, \quad \dots, \quad a_{KK} \neq 0.$$

Then the vectors  $\check{M}_1^*, \dots, \check{M}_K^*$  are also linearly independent."

In fact, if there were a relation of the form

$$\lambda_1 \check{M}_1^* + \dots + \lambda_K \check{M}_K^* = \sigma$$

and the factors  $\lambda_1, \dots, \lambda_K$  were not all equal to zero, then there would be a last factor differing from zero, say  $\lambda_L$ , and we should have

$$\lambda_1 \check{M}_1^* + \dots + \lambda_L \check{M}_L^* = \sigma \quad (\lambda_L \neq 0)$$

if we were to replace  $\check{M}_1^*, \dots, \check{M}_L^*$  by the expressions (2), we should get a relation of the form

$$\mu_1 \check{M}_1 + \dots + \mu_{L-1} \check{M}_{L-1} + a_{LL} \lambda_L \check{M}_L = \sigma,$$

which is impossible on account of  $a_{LL} \neq 0$ ,  $\lambda_L \neq 0$  and the presupposed linear independency of  $\check{M}_1, \dots, \check{M}_K$ .

In order to prove some further theorems it is convenient—but not necessary—to make use of the following fundamental

theorem concerning systems of homogeneous linear equations.

*"A necessary and sufficient condition that the system of equations*

$$a_{11} x_1 + \dots + a_{1n} x_n = 0$$

$$-----$$

$$a_{n1} x_1 + \dots + a_{nn} x_n = 0$$

*should have no other solution than  $x_1 = x_2 = \dots = x_n = 0$  is*

$$\begin{vmatrix} a_{11} & \dots & a_{1n} \\ a_{n1} & \dots & a_{nn} \end{vmatrix} \neq 0."$$

From this statement at once follows:

**THEOREM 2.** *"A necessary and sufficient condition that the vectors*

$$\check{M}_1 = (u_{11}, u_{12}, \dots, u_{1n})$$

$$-----$$

$$\check{M}_n = (u_{n1}, u_{n2}, \dots, u_{nn})$$

*should be linearly independent is*

$$\begin{vmatrix} u_{11} & \dots & u_{1n} \\ u_{n1} & \dots & u_{nn} \end{vmatrix} \neq 0."$$

In fact, linear dependence of  $\check{M}_1, \dots, \check{M}_n$  is equivalent to the existence of values  $\lambda_1, \dots, \lambda_n$ , not all equal to zero, satisfying

$$\lambda_1 u_{11} + \lambda_2 u_{21} + \dots + \lambda_n u_{n1} = 0$$

$$\lambda_1 u_{1n} + \lambda_2 u_{2n} + \dots + \lambda_n u_{nn} = 0,$$

and the determinant of these equations is equal to the determinant above.

**THEOREM 3.** *"If  $\check{M}_1, \dots, \check{M}_k$  are linearly independent, the number  $k$  of the vectors cannot exceed the number  $n$  of the components:  $k \leq n$ ."*

We prove this theorem by showing:

"If  $n+1$  vectors

$$\check{u}_1 = (u_{11}, \dots, u_{1n})$$

-----

$$\check{u}_{n+1} = (u_{n+1,1}, \dots, u_{n+1,n})$$

are given, they are linearly dependent."

For obviously, the determinant of the equations

$$\lambda_1 u_{11} + \dots + \lambda_{n+1} u_{n+1,1} = 0$$

-----

$$\lambda_1 u_{1n} + \dots + \lambda_{n+1} u_{n+1,n} = 0$$

$$\lambda_1 0 + \dots + \lambda_{n+1} 0 = 0$$

vanishes, hence the system possesses a solution  $\lambda_1, \dots, \lambda_{n+1}$  different from  $0, \dots, 0$ , and with such values  $\lambda_1, \dots, \lambda_{n+1}$  the first  $n$  equations mean

$$\lambda_1 \check{u}_1 + \dots + \lambda_{n+1} \check{u}_{n+1} = 0.$$

## 6. NORMALIZED ORTHOGONAL SYSTEMS OF VECTORS

If  $\check{u}^2 = 1$ , the vector  $\check{u}$  is said to be *normalized*. Every vector  $\check{u} \neq 0$  can be normalized by multiplying it by  $\frac{1}{\sqrt{\check{u}^2}}$ .

If  $\check{u}\check{v} = 0$ , the pair of vectors  $\check{u}$  and  $\check{v}$  is said to be *orthogonal*. The vectors  $\check{u}_1, \dots, \check{u}_k$  are said to form an *orthogonal system*, if every pair of them is orthogonal.

Finally the vectors  $\check{u}_1, \dots, \check{u}_k$  are said to form a *normalized orthogonal system*, if they form an orthogonal system and each of them is normalized. Accordingly a normalized orthogonal system is characterized by the conditions

$$(3) \quad \check{u}_p \check{u}_q = \begin{cases} 1 & \text{for } p = q \\ 0 & \text{for } p \neq q. \end{cases}$$

Vectors forming a normalized orthogonal system necessarily are linearly independent. For, from

$$\lambda_1 \check{u}_1 + \dots + \lambda_k \check{u}_k = 0$$

follows

$$(\lambda_1 \check{u}_1 + \dots + \lambda_K \check{u}_K) \check{u}_\pi = \sigma \check{u}_\pi = 0 \quad (\pi = 1, 2, \dots, K)$$

or

$$\lambda_1 \check{u}_1 \check{u}_\pi + \dots + \lambda_K \check{u}_K \check{u}_\pi = 0,$$

and from (3):

$$\lambda_\pi = 0 \quad (\pi = 1, 2, \dots, K).$$

## 7. THE PROCESS OF ORTHOGONALIZATION

THEOREM 4. "If the vectors  $\check{u}_1, \dots, \check{u}_K$  are linearly independent, there is one and only one scheme of values

$$\begin{array}{ll} \beta_{11} & (\beta_{11} > 0) \\ \beta_{21} & \beta_{22} & (\beta_{22} > 0) \\ \dots & \dots & \dots \\ \beta_{K1} & \beta_{K2} & \dots & \beta_{KK} & (\beta_{KK} > 0) \end{array}$$

so that the vectors

$$\begin{aligned} \check{u}_1 &= \beta_{11} \check{u}_1 \\ \check{u}_2 &= \beta_{21} \check{u}_1 + \beta_{22} \check{u}_2 \\ &\dots \dots \dots \\ \check{u}_K &= \beta_{K1} \check{u}_1 + \beta_{K2} \check{u}_2 + \dots + \beta_{KK} \check{u}_K \end{aligned} \quad (4)$$

form a normalized orthogonal system."

To prove this theorem, fundamental for our analysis, let us consider

$$\begin{aligned} \check{u}_1 &= \check{u}_1 \\ \check{u}_2 &= \gamma_{21} \check{u}_1 + \check{u}_2 \\ &\dots \dots \dots \\ \check{u}_K &= \gamma_{K1} \check{u}_1 + \gamma_{K2} \check{u}_2 + \dots + \check{u}_K \end{aligned} \quad (5)$$

From theorem 1 it follows that the vectors  $\check{u}_1, \dots, \check{u}_K$  are linearly independent. — Let us assume we have already proved that there is one and only one system of values  $\gamma$  so that (5) is an orthogonal system. Then it follows firstly that the coefficients  $\beta$  in (4) can be chosen in at least one suitable manner. For we have

$$\lambda_1 = \sqrt{\check{u}_1^2} > 0, \dots, \lambda_K = \sqrt{\check{u}_K^2} > 0,$$

and

$$\beta_{11} = \frac{1}{\lambda}; \beta_{21} = \frac{\gamma_{21}}{\lambda}, \beta_{22} = \frac{1}{\lambda}; \dots; \beta_{K1} = \frac{\gamma_{K1}}{\lambda}, \dots, \beta_{KK} = \frac{1}{\lambda}$$

are suitable values. Secondly we can deduce the *uniqueness* of the coefficients  $\beta$  in (4). For suppose  $\beta$  and  $\beta^*$  to be two suitable systems of coefficients; this suggests that we form

$$\gamma_{21} = \frac{\beta_{21}}{\beta_{22}}; \gamma_{31} = \frac{\beta_{31}}{\beta_{33}}, \gamma_{32} = \frac{\beta_{32}}{\beta_{33}}; \dots; \gamma_{k1} = \frac{\beta_{k1}}{\beta_{kk}}, \dots; \gamma_{k; k-1} = \frac{\beta_{k; k-1}}{\beta_{kk}},$$

associated with

$$AD_1 = \frac{1}{\beta_{11}} AD_1, \dots, AD_k = \frac{1}{\beta_{kk}} AD_k,$$

and

$$\gamma_{21}^* = \frac{\beta_{21}^*}{\beta_{22}^*}; \gamma_{31}^* = \frac{\beta_{31}^*}{\beta_{33}^*}, \gamma_{32}^* = \frac{\beta_{32}^*}{\beta_{33}^*}; \dots; \gamma_{k1}^* = \frac{\beta_{k1}^*}{\beta_{kk}^*}, \dots; \gamma_{k; k-1}^* = \frac{\beta_{k; k-1}^*}{\beta_{kk}^*},$$

associated with

$$AD_1^* = \frac{1}{\beta_{11}^*} AD_1^*, \dots, AD_k^* = \frac{1}{\beta_{kk}^*} AD_k^*.$$

The vectors  $AD_1, \dots, AD_k$  as well as  $AD_1^*, \dots, AD_k^*$  form orthogonal systems of the type (2), hence

$$AD_1^* = AD_1, \dots, AD_k^* = AD_k$$

and furthermore

$$AD_1^* = AD_1, \dots, AD_k^* = AD_k.$$

Finally, because of the linear independence of  $\check{M}_1, \dots, \check{M}_k$ :

$$\beta_{11}^* = \beta_{11}; \beta_{21}^* = \beta_{21}, \beta_{22}^* = \beta_{22}; \dots; \beta_{k1}^* = \beta_{k1}, \beta_{kk}^* = \beta_{kk}.$$

Accordingly we may confine ourselves to proving the existence and uniqueness of suitable coefficients  $\gamma$  in (5).

This proposition is true for  $k=1$ . Let  $k \geq 2$  arbitrarily, and assume the proposition to be proved up to  $k-1$ . The vectors  $AD_1, \dots, AD_{k-1}$  therefore are orthogonal, and we have to show only: There is one and only one set of values  $\gamma_{k1}, \dots, \gamma_{k; k-1}$  so that the conditions

$$(6) \quad AD_1 AD_k = 0, \quad AD_2 AD_k = 0, \dots, AD_{k-1} AD_k = 0$$

are satisfied.

The vectors  $\check{M}_1, \dots, \check{M}_{k-1}$  can be represented as linear combinations of  $AD_1, \dots, AD_{k-1}$ :

$$\check{M}_1 = AD_1$$

$$\check{M}_2 = c_{21} AD_1 + AD_2$$

$$\check{M}_{k-1} = c_{k-1,1} AD_1 + \dots + c_{k-1, k-2} AD_{k-2} + AD_{k-1}.$$

We introduce this into  $AD_K$ , and get

$$(7) \quad AD_K = C_{K1} AD_1 + \dots + C_{K,K-1} AD_{K-1} + \check{M}_K$$

with

$$(8) \quad \begin{cases} C_{K1} = \check{\gamma}_{K1} + c_{21} \check{\gamma}_{K2} + \dots + c_{K-1,1} \check{\gamma}_{K,K-1} \\ C_{K2} = \dots \quad \check{\gamma}_{K2} + \dots + c_{K-1,2} \check{\gamma}_{K,K-1} \\ \dots \dots \dots \\ C_{K,K-1} = \dots \quad \check{\gamma}_{K,K-1} \end{cases}$$

From the linear independence of  $AD_1, \dots, AD_{K-1}$  we have  $AD_1^2 > 0, \dots, AD_{K-1}^2 > 0$ , and therefore we can deduce from (7):

$$(9) \quad C_{K1} = - \frac{\check{M}_K AD_1}{\sqrt{AD_1^2}}, \dots, C_{K,K-1} = - \frac{\check{M}_K AD_{K-1}}{\sqrt{AD_{K-1}^2}}.$$

The coefficients  $\check{\gamma}_{K1}, \dots, \check{\gamma}_{K,K-1}$  having to satisfy the equations (8) with the values (9) of  $C_{K1}, \dots, C_{K,K-1}$ , there can exist *only one* suitable system  $\check{\gamma}_{K1}, \dots, \check{\gamma}_{K,K-1}$ .

*Conversely*: if  $C_{K1}, \dots, C_{K,K-1}$  are chosen according to (9), and then  $\check{\gamma}_{K1}, \dots, \check{\gamma}_{K,K-1}$  calculated from (8), evidently there results a vector  $AD_K$  satisfying (6).

We add:

**THEOREM 5.** "If  $\check{M}_1, \dots, \check{M}_K$  are linearly independent, and  $AD_1, \dots, AD_K$  is the corresponding normalized orthogonal system, then the normalized orthogonal system  $AD_1^*, \dots, AD_K^*$  corresponding to

$$\check{M}_1^* = a_{11} \check{M}_1 \quad (a_{11} > 0)$$

$$\check{M}_2^* = a_{21} \check{M}_1 + a_{22} \check{M}_2 \quad (a_{22} > 0)$$

$$\dots \dots \dots$$

$$\check{M}_K^* = a_{K1} \check{M}_1 + a_{K2} \check{M}_2 + \dots + a_{KK} \check{M}_K \quad (a_{KK} > 0)$$

is identical with  $AD_1, \dots, AD_K$ ."

Obviously the vectors  $AD^*$  are of the form

$$AD_1^* = B_{11} \check{M}_1 \quad (B_{11} > 0)$$

$$AD_2^* = B_{21} \check{M}_1 + B_{22} \check{M}_2 \quad (B_{22} > 0)$$

$$\dots \dots \dots$$

$$AD_K^* = B_{K1} \check{M}_1 + B_{K2} \check{M}_2 + \dots + B_{KK} \check{M}_K \quad (B_{KK} > 0)$$

The proof of theorem 5 now follows as an immediate application of theorem 4.

#### 8. COMPLETE SYSTEMS OF NORMALIZED ORTHOGONAL VECTORS

A system of normalized orthogonal vectors  $MO_1, \dots, MO_K$  is said to be *complete* if, corresponding to every arbitrary vector  $\check{M}$ , there exist coefficients  $c_1, \dots, c_K$  so that

$$(10) \quad \{ \check{M} - (c_1 MO_1 + \dots + c_K MO_K) \}^2 = 0$$

holds. Evidently, (10) is equivalent to

$$\check{M} = c_1 MO_1 + \dots + c_K MO_K.$$

THEOREM 6. "If the vectors  $MO_1, \dots, MO_n$  ( $K=n$ ) form a normalized orthogonal system, then this system is COMPLETE."

Proof. According to theorem 3, the  $n+1$  vectors  $MO_1, \dots, MO_n, \check{M}$  are linearly dependent, i.e. there is an equality

$$\lambda_1 MO_1 + \dots + \lambda_n MO_n + \lambda \check{M} = 0,$$

and  $\lambda_1, \dots, \lambda_n, \lambda$  are not all equal to zero. The vectors  $MO_1, \dots, MO_n$ , being linearly independent, we have necessarily  $\lambda \neq 0$ . Hence

$$\check{M} = -\frac{\lambda_1}{\lambda} MO_1 - \dots - \frac{\lambda_n}{\lambda} MO_n.$$

The condition  $K=n$  is also *necessary* for completeness, but we shall not have to make use of it.

#### 9. APPROXIMATION IN THE MEAN

Let us consider a normalized orthogonal system  $MO_1, \dots, MO_K$ , and an arbitrary vector  $\check{M}$ . We wish to determine the coefficients  $b_1, \dots, b_K$  in such a way that

$$\left\{ \check{M} - \sum_{k=1}^K b_k MO_k \right\}^2$$

assumes a minimum. If there exists a suitable set of coefficients, we say that the corresponding linear combination  $b_1 MO_1 + \dots + b_K MO_K$  gives a "*best approximation in the mean*" to the vector  $\check{M}$ .



The following transformations will at once clear up the situation:

$$\begin{aligned} \left\{ \check{M} - \sum_{x=1}^k b_x m_x \right\}^2 &= \check{M}^2 - 2 \sum_{x=1}^k b_x m_x \check{M} + \sum_{s,s=1}^k b_s b_s m_s m_s \\ &= \check{M}^2 - \sum_{x=1}^k (m_x \check{M})^2 + \sum_{x=1}^k (m_x \check{M})^2 - 2 \sum_{x=1}^k b_x m_x \check{M} + \sum_{x=1}^k b_x^2 \\ &= \check{M}^2 - \sum_{x=1}^k (m_x \check{M})^2 + \sum_{x=1}^k \{ b_x - m_x \check{M} \}^2, \end{aligned}$$

and if we designate

$$a_1 = m_1 \check{M}, \quad a_2 = m_2 \check{M}, \dots, \quad a_k = m_k \check{M}$$

we have the fundamental equation

$$(11) \quad \left\{ \check{M} - \sum_{x=1}^k b_x m_x \right\}^2 = \check{M}^2 - \sum_{x=1}^k a_x^2 + \sum_{x=1}^k (b_x - a_x)^2.$$

On the right hand, the coefficients  $b_1, \dots, b_k$  are not met with but in the last sum, and this sum assumes a minimum for  $b_x = a_x$  only. By that, we have:

**THEOREM 7.** "Among all linear combinations of the normalized orthogonal vectors  $m_1, \dots, m_k$  there is one and only one which gives a best approximation in the mean to the vector  $\check{M}$ , and the 'best coefficients' are  $a_1, a_2, \dots, a_k$ ."

The equation (11) admits some important conclusions concerning the coefficients  $a_1, \dots, a_k$ . By putting

$$b_1 = a_1, \dots, b_k = a_k$$

we derive

$$(12) \quad \left\{ \check{M} - \sum_{x=1}^k a_x m_x \right\}^2 = \check{M}^2 - \sum_{x=1}^k a_x^2.$$

The left side herein evidently is not negative, hence

$$(13) \quad a_1^2 + a_2^2 + \dots + a_k^2 \leq \check{M}^2.$$

Finally, if  $m_1, \dots, m_n$  is a COMPLETE system of normalized orthogonal vectors, the preceding reasonings of course hold for every  $k = 1, 2, \dots, n$ . But we can show more than (13), viz.

$$(14) \quad a_1^2 + a_2^2 + \dots + a_n^2 = \check{M}^2.$$

Indeed, according to the definition of completeness we have with suitable  $t_1, \dots, t_n$ :

$$\left\{ \tilde{u} - \sum_{\nu=1}^n t_{\nu} u_{\nu} \right\}^2 = 0$$

and a fortiori, by theorem 7,

$$\left\{ \tilde{u} - \sum_{\nu=1}^n a_{\nu} u_{\nu} \right\}^2 = 0,$$

which is, regarding (12), equivalent to (14).

#### 10. THE TCHEBYCHEF COEFFICIENTS

Let  $\xi_1 < \xi_2 < \dots < \xi_n$

be a set of best values corresponding to a given frequency function  $\varphi(x)$  (see No. 2). We form

$$(15) \quad \begin{aligned} \mathcal{E}_0 &= (1, 1, \dots, 1) \\ \mathcal{E}_1 &= (\xi_1, \xi_2, \dots, \xi_n) \\ \mathcal{E}_2 &= (\xi_1^2, \xi_2^2, \dots, \xi_n^2) \\ &\dots\dots\dots \\ \mathcal{E}_{n-1} &= (\xi_1^{n-1}, \xi_2^{n-1}, \dots, \xi_n^{n-1}) \end{aligned}$$

The vectors  $\mathcal{E}_0, \dots, \mathcal{E}_{n-1}$  are linearly independent. For

means  $\lambda_0 \mathcal{E}_0 + \dots + \lambda_{n-1} \mathcal{E}_{n-1} = 0$   
 $\lambda_0 + \lambda_1 \xi_{\nu} + \lambda_2 \xi_{\nu}^2 + \dots + \lambda_{n-1} \xi_{\nu}^{n-1} = 0 \quad (\nu=1, \dots, n),$   
 that is, the polynomial

$P(x) = \lambda_0 + \lambda_1 x + \dots + \lambda_{n-1} x^{n-1}$   
 of degree  $\leq (n-1)$  possesses  $n$  different zeros  $\xi_1, \dots, \xi_n$ .  
 But the number of zeros of a polynomial cannot exceed its degree unless all coefficients vanish. Hence  $\lambda_0 = \lambda_1 = \dots = \lambda_{n-1} = 0$ .

Let us designate the (complete) set of normalized orthogonal vectors corresponding to  $\mathcal{E}_0, \dots, \mathcal{E}_{n-1}$  by  $\mathcal{F}_0, \mathcal{F}_1, \dots, \mathcal{F}_{n-1}$ .

When we have to deal with a set of observations  $x_1, \dots, x_n$ , there will not be any practical loss of generality if we assume these values arranged according to magnitude,

$$x_1 \leq x_2 \leq \dots \leq x_n,$$

and to be *not all equal*. Then we define the vector  $\mathcal{E}$  by

$$\mathcal{E} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n),$$

and we propose to call the coefficients

$$a_0 = \mathcal{J}_0 \mathcal{E}, \quad a_1 = \mathcal{J}_1 \mathcal{E}, \quad \dots, \quad a_{n-1} = \mathcal{J}_{n-1} \mathcal{E}$$

"Tchebychef Coefficients" of  $\mathcal{E}$  ..

The central position of the Tchebychef coefficients for analyzing purposes is pointed out by the following theorems 8 and 9.

THEOREM 8. "The set  $\mathcal{J}_0, \mathcal{J}_1, \dots, \mathcal{J}_{n-1}$  and a fortiori the Tchebychef coefficients  $a_0, a_1, \dots, a_{n-1}$  of the observations  $x_1, x_2, \dots, x_n$  do not depend on the special frequency function  $\varphi(x)$ , but on the type only to which  $\varphi(x)$  belongs."

To prove this theorem, let us consider, besides  $\varphi(x)$ , an arbitrary individual of its type,

$$\tilde{\varphi}(x) = \frac{1}{\alpha} \varphi\left(\frac{1}{\alpha}(x-\beta)\right) \quad (\alpha > 0).$$

The best values corresponding to  $\tilde{\varphi}(x)$  are (see No. 2)

$$\tilde{\xi}_1 = \alpha \cdot \xi_1 + \beta, \dots, \quad \tilde{\xi}_n = \alpha \cdot \xi_n + \beta,$$

and we deduce, if  $\tilde{\mathcal{E}}_0, \dots, \tilde{\mathcal{E}}_{n-1}$  designate the vectors (15) obtained from  $\tilde{\xi}_1, \dots, \tilde{\xi}_n$  instead of  $\xi_1, \dots, \xi_n$ ,

$$\tilde{\mathcal{E}}_0 = \mathcal{E}_0$$

$$\tilde{\mathcal{E}}_1 = \beta \mathcal{E}_0 + \alpha \mathcal{E}_1$$

$$\tilde{\mathcal{E}}_2 = \beta^2 \mathcal{E}_0 + 2\beta\alpha \mathcal{E}_1 + \alpha^2 \mathcal{E}_2$$

$$\dots \dots \dots$$

$$\tilde{\mathcal{E}}_{n-1} = \beta^{n-1} \mathcal{E}_0 + (n-1)\beta^{n-2}\alpha \mathcal{E}_1 + \dots + \alpha^{n-1} \mathcal{E}_{n-1}.$$

Hence, by an application of theorem 5, the normalized orthogonal vectors  $\tilde{\mathcal{J}}_0, \dots, \tilde{\mathcal{J}}_{n-1}$  are identical with  $\mathcal{J}_0, \dots, \mathcal{J}_{n-1}$ .

If we choose a new unit of measurement and a new origin, that is to say if we perform a transformation

$$x_\nu^* = \alpha x_\nu + \beta, \quad \xi_\nu^* = \alpha \xi_\nu + \beta \quad (\alpha > 0),$$

the vectors  $\mathcal{J}_0, \dots, \mathcal{J}_{n-1}$  do not change (by the reasoning just finished). The vector  $\mathcal{E}$  changes into

$$\mathcal{E}^* = \alpha \mathcal{E} + \beta \mathcal{N} \quad (\mathcal{N} = (1, 1, \dots, 1)),$$

and we have:

\*THEOREM 9. "If a new unit of measurement and a new origin are introduced, say

$$x = \gamma x + \beta \quad (\gamma > 0),$$

then the Tchebycheff coefficients change into

$$a_0^* = \gamma a_0 + \beta; \quad a_1^* = \gamma a_1, \quad a_2^* = \gamma a_2, \dots, a_{n-1}^* = \gamma a_{n-1}."$$

## 11. MEAN AND DISPERSION. COEFFICIENTS

### OF SKEWNESS AND KURTOSIS

Preparatory to the definition in this chapter, let us consider  $a_0$  and  $a_1$  especially. To begin with, we have

$$j_0 = \bar{x}_0 = (1, 1, \dots, 1)$$

and therefore

$$a_0 = \frac{1}{n} (x_1 + x_2 + \dots + x_n).$$

The proof of theorem 4 furnishes a convenient way to compute  $j_1$ . We put

$$y_0 = \bar{x}_0; \quad y_1 = \gamma \bar{x}_0 + \bar{x}_1$$

and determine  $\gamma$  so that  $y_0 y_1 = 0$ :

$$\gamma = - \frac{\bar{x}_0 \bar{x}_1}{\bar{x}_0^2} = - \frac{1}{n} (\xi_1 + \dots + \xi_n).$$

With the designations

$$m_1 = \frac{1}{n} (\xi_1 + \dots + \xi_n) \quad m_2 = \frac{1}{n} (\xi_1^2 + \dots + \xi_n^2)$$

we obtain

$$\bar{x}_1 \bar{x}_0 = m_1, \quad \bar{x}_1^2 = m_2.$$

Hence

$$\begin{aligned} y_1 &= -m_1 \bar{x}_0 + \bar{x}_1, & y_1^2 &= m_2 - m_1^2 \\ j_1 &= \frac{1}{\sqrt{y_1^2}} y_1 = \frac{1}{\sqrt{m_2 - m_1^2}} (-m_1 \bar{x}_0 + \bar{x}_1) \\ &= \frac{1}{\sqrt{m_2 - m_1^2}} (\xi_1 - m_1, \xi_2 - m_1, \dots, \xi_n - m_1) \\ a_1 &= j_1 \bar{x}_0 = \frac{1}{\sqrt{m_2 - m_1^2}} (-m_1 \bar{x}_0 \bar{x}_0 + \bar{x}_1 \bar{x}_0) \\ &= \frac{\frac{1}{n} (\xi_1 x_1 + \dots + \xi_n x_n) - m_1 \cdot \frac{1}{n} (x_1 + \dots + x_n)}{\sqrt{m_2 - m_1^2}}. \end{aligned}$$

Concerning  $a_1$  we have now to deal with a theorem which is of the greatest importance for our purposes.

THEOREM 10. "The Tchebycheff coefficient  $a_1$  is always positive:

$$a_1 > 0."$$

For a proof we can proceed as follows: if we designate the components of  $g_1$  by  $S_1, \dots, S_n$ , we have

$$S_1 < S_2 < \dots < S_n \quad \text{and} \quad S_1 + \dots + S_n = 0.$$

From this we deduce the existence of a subscript  $\nu$  so that

$$S_1 < \dots < S_\nu < 0 \leq S_{\nu+1} < \dots < S_n.$$

Let us put

$$z_1 = S_1, \quad z_2 = S_1 + S_2, \quad \dots, \quad z_n = S_1 + \dots + S_n.$$

Then we have

$$z_1 < 0, \dots, z_\nu < 0; \quad z_{\nu+1} < z_{\nu+2} < \dots < z_n = 0,$$

which gives

$$z_1 < 0, \quad z_2 < 0, \quad \dots, \quad z_{n-1} < 0.$$

On the other hand, the identity

$$S_1 x_1 + \dots + S_n x_n = -z_1 (x_2 - x_1) - z_2 (x_3 - x_2) - \dots - z_{n-1} (x_n - x_{n-1})$$

holds. The differences  $x_2 - x_1, \dots, x_n - x_{n-1}$  are all  $\geq 0$ , and  $x_1, \dots, x_n$  being subjected to the condition not to be all equal, at least one difference really is positive. Hence

$$a_1 = \frac{1}{n^2} (S_1 x_1 + \dots + S_n x_n) > 0.$$

There are no restrictions for the Tchebycheff coefficients different from  $a_1$ , as far as their signs are concerned.

The reader, after having verified the truth of the following statement, will now be prepared to accept the definition below.

"If the vector  $\mathcal{C}$  is of the form

$$\mathcal{C} = \ell_0 \mathcal{C}_0 + \ell_1 \mathcal{C}_1 + \ell_2 \mathcal{C}_2,$$

the sign of  $a_2$  coincides with that of  $\ell_2$ ; if it is of the form

$$\mathcal{C} = \ell_0 \mathcal{C}_0 + \ell_1 \mathcal{C}_1 + \ell_2 \mathcal{C}_2 + \ell_3 \mathcal{C}_3$$

the sign of  $a_3$  coincides with that of  $\ell_3$ ; and so on."

DEFINITION. "A type of frequency function being given, the Tchebycheff coefficients  $a_0, a_1, a_2, a_3$  of the observations  $x_1, x_2, \dots, x_n$  shall be called:

$a_0 = M$  = MEAN of the Observations

$a_1 = \sigma$  = DISPERSION of the Observations

$a_2^*$  = TCHEBYCHEF COEFFICIENT OF SKEWNESS of the Observations

$a_3$  = TCHEBYCHEF COEFFICIENT OF KURTOSIS of the Observations."

We do not believe the Tchebychef coefficients with a higher subscript than 3 to be of any practical interest.

## 12. MEASURES OF SKEWNESS AND KURTOSIS

No matter how the mean and the dispersion of a set of observations are defined, the dispersion will always have to depend on the unit of measurement, and the mean furthermore on the origin. But the case is a different one concerning the concepts of skewness and kurtosis. Here it is reasonable to raise the question for measures in the strict sense. It is obvious that such measures will be obtained if the set of observations is—by a convenient choice of a new unit—brought to the dispersion 1; the new Tchebychef coefficients of skewness and kurtosis will be suitable. This leads to the

DEFINITION. "With the designations of the preceding chapter, the ratios  $\frac{a_2}{a_1}$  and  $\frac{a_3}{a_1}$  shall be called:

$\frac{a_2}{a_1} = S$  = MEASURE OF SKEWNESS of the Observations

$\frac{a_3}{a_1} = K$  = MEASURE OF KURTOSIS of the Observations."

There will be no misunderstanding if we use the words "Skewness" and "Kurtosis" instead of "Measure of Skewness" and "Measure of Kurtosis".—Utilizing theorems 8 and 9, we have at once:

THEOREM 11. "The measures of skewness and kurtosis depend on the type of frequency function and on the observations  $\mathcal{E}$  only; they are independent of origin and unit of measurement."

## 13. MEANING OF SKEWNESS AND KURTOSIS

To secure an idea of the mechanism of skewness and kurtosis, let us construct some examples which show these phenomena in

complete purity. We will use the step function, and we intend to choose the values  $x_2, \dots, x_n$  so that they are affected—apart from the inevitable dispersion—in the first place with skewness only, in the second place with kurtosis only.

We take  $n = 10$ , and for the convenience of the reader we actually write down the vectors  $\beta_0, \dots, \beta_3$ . We observe however that in practice one will never evaluate these vectors, but rather compute the Tchebychef coefficients in the direct manner described in No. 15.

We obtain

$\nu$	$\beta_0$	$\beta_1$	$\beta_2$	$\beta_3$
1	1	-1.56670	+ 1.65145	- 1.43388
2	1	-1.21855	+ .55048	+ .47796
3	1	-.87039	- .27524	+ 1.19490
4	1	-.52223	-.82572	+ 1.05834
5	1	-.17408	- 1.10096	+ .40968
6	1	+ .17408	- 1.10096	- .40968
7	1	+ .52223	- .82572	- 1.05834
8	1	+ .87039	- .27524	- 1.19490
9	1	+ 1.21855	+ .55048	- .47796
10	1	+ 1.56670	+ 1.65145	+ 1.43388

We shall have to come back to these vectors in No. 17. For this reason they have been calculated more accurately than is necessary here.

1 a. Positive skewness.

$$\mathcal{E} = \beta_1 + \frac{1}{5} \beta_2 \quad (a_1 = 1, a_2 = +\frac{1}{5}; a_\nu = 0 \text{ otherwise}).$$

1 b. Negative skewness.

$$\mathcal{E} = \beta_1 - \frac{1}{5} \beta_2 \quad (a_1 = 1, a_2 = -\frac{1}{5}; a_\nu = 0 \text{ otherwise}).$$

2 a. Positive kurtosis.

$$\mathcal{E} = \beta_1 + \frac{1}{10} \beta_3 \quad (a_1 = 1, a_3 = +\frac{1}{10}; a_\nu = 0 \text{ otherwise}).$$

2 b. Negative kurtosis.

$$\mathcal{E} = \beta_1 - \frac{1}{10} \beta_3 \quad (a_1 = 1, a_3 = -\frac{1}{10}; a_\nu = 0 \text{ otherwise}).$$

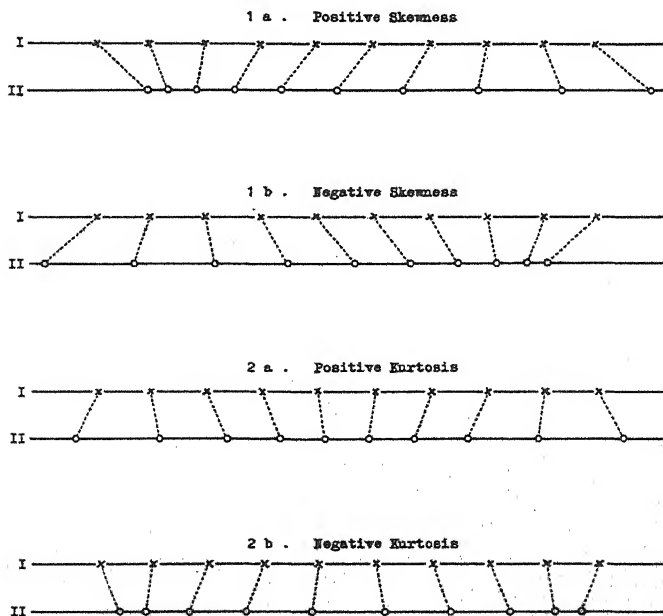
The components of the different vectors  $\mathcal{E}$  are put together in the table below:

$\nu$	$\bar{z}_1 + \frac{1}{5} \bar{z}_2$	$\bar{z}_1 - \frac{1}{5} \bar{z}_2$	$\bar{z}_1 + \frac{1}{10} \bar{z}_3$	$\bar{z}_1 - \frac{1}{10} \bar{z}_3$
1	- 1.236	- 1.897	- 1.710	- 1.423
2	- 1.108	- 1.329	- 1.171	- 1.266
3	- .925	- .815	- .751	- .990
4	- .687	- .357	- .416	- .628
5	- .394	+ .046	- .133	- .215
6	- .046	+ .394	+ .133	+ .215
7	+ .357	+ .687	+ .416	+ .628
8	+ .815	+ .925	+ .751	+ .990
9	+ 1.329	+ 1.108	+ 1.171	+ 1.266
10	+ 1.897	+ 1.236	+ 1.710	+ 1.423

To illustrate the preceding, we compare the vectors  $\mathcal{E}$  with their corresponding "best systems of best values", that is to say with the vectors

$$\bar{\mathcal{E}} = a_0 \bar{z}_0 + a_1 \bar{z}_1,$$

and carry it through with some figures. We place the components of  $\bar{\mathcal{E}}$  on a horizontal straight line I, the components of  $\mathcal{E}$  on a second straight line II below:





The reader should settle his mind upon the fact that the general behaviour of observations affected with skewness only or kurtosis only is always the same, no matter which type of frequency function is considered.—The meaning of skewness and kurtosis can be, generally speaking, expressed by:

*Positive Skewness = Overconcentration to the Left*

*Negative Skewness = Overconcentration to the Right*

*Positive Kurtosis = Overconcentration near the Mean*

*Negative Kurtosis = Underconcentration near the Mean.*

#### 14. MEASURES OF APPROXIMATION

Let  $\mathcal{T}$  be a type of frequency function,  $\mathcal{E} = (x_1, \dots, x_n)$  a set of observations, and  $k \geq 1$  a "degree of approximation", that is the subscript in the sum  $a_0 z_0 + \dots + a_k z_k$ . The expression

$$\{\mathcal{E} - (a_0 z_0 + \dots + a_k z_k)\}^2$$

will give us a clue to the quality of approximation to the vector  $\mathcal{E}$  which is obtained on the basis of the type  $\mathcal{T}$  and the degree of approximation  $k$ . But the expression above of course is not yet fit to be taken as a measure of the quality of approximation. Therefore it will be necessary in the first instance to modify it so that it will become not only independent of the origin, but also independent of the unit of measurement.

Regarding theorem 9, and making reflections customary in situations of this kind, we are almost compulsorily led to the

DEFINITION. "The values

$$M_k = \sqrt{\frac{a_1^2 + \dots + a_k^2}{\mathcal{E}^2 - a_0^2}} \quad (k = 1, 2, \dots, n-1)$$

shall be called MEASURES OF APPROXIMATION OF THE DEGREES  $k$ ."

THEOREM 12. "The measures of approximation  $M_k$  depend on the type  $\mathcal{T}$  and on the observations  $\mathcal{E}$  only; they are independent of origin and unit of measurement. Furthermore they satisfy

$$0 < M_1 \leq M_2 \leq \dots \leq M_{n-1} = 1."$$

All is clear if we write—utilizing the relation (14) in No. 9—

$M_K$  in the form

$$M_K^2 = \frac{a_1^2 + a_2^2 + \dots + a_K^2}{a_1^2 + a_2^2 + \dots + a_{n-1}^2},$$

paying attention to theorems 6, 8 and 9.

If  $M_K$  is not much smaller than 1, the approximation of degree  $K$  will be estimated to be good. If  $T$  and  $T^*$  are two types of frequency function,  $M_K$  and  $M_K^*$  the corresponding measures of approximation, and if  $M_K^* \geq M_K$ , we say:  $T^*$  is, for the degree  $K$ , better than  $T$  (equivalence not excluded). If  $M_1^* \geq M_1, \dots, M_K^* \geq M_K$ , we say:  $T^*$  is, up to the degree  $K$ , better than  $T$ .

Clearly we may base upon these concepts a method of curve-fitting. A full account will be given in a future note.

#### 15. COMPUTATION OF THE TCHEBYCHEF COEFFICIENTS

If the vectors  $z_0, \dots, z_K$  are already known, the finding of  $a_0, \dots, a_K$  is, according to their definition, very simple. But the actual calculation of  $z_0, \dots, z_K$  is embarrassing, especially if  $n$  is large. We already mentioned that this can be and should be avoided, and we recommend the following procedure.

We form, just as in the proof of theorem 4,

$$\begin{aligned} y_0 &= e_0 \\ y_1 &= \gamma_{10} e_0 + e_1 \\ &\dots \\ y_K &= \gamma_{K0} e_0 + \dots + \gamma_{K, K-1} e_{K-1} + e_K, \end{aligned}$$

and to determine the coefficients  $\gamma$ , we demand that the vectors  $y$  be orthogonal. Let  $x$  be an arbitrary subscript among  $0, 1, \dots, K$ . Then at least it must be true that

$$(16) \quad y_x y_0 = 0, \dots, y_x y_{x-1} = 0,$$

and a fortiori

$$y_x (c_0 y_0 + \dots + c_{x-1} y_{x-1}) = 0$$

for arbitrary values  $c_0, \dots, c_{x-1}$ . But  $e_0, \dots, e_{x-1}$  are linear

combinations of  $\mathcal{Y}_0, \dots, \mathcal{Y}_{x-1}$ , hence

$$(17) \quad \mathcal{Y}_x \mathcal{E}_0 = 0, \dots, \mathcal{Y}_x \mathcal{E}_{x-1} = 0.$$

For abbreviation let us designate the moments of the best values  $\xi_1, \dots, \xi_n$  by

$$m_r = \frac{1}{n} (\xi_1^r + \xi_2^r + \dots + \xi_n^r) \quad (r = 0, 1, \dots).$$

Obviously we have

$$\mathcal{E}_p \mathcal{E}_q = m_{p+q} \quad (p, q = 0, 1, \dots),$$

and the equation (17) produces

$$(18) \quad \begin{cases} m_0 \gamma_{x0} + m_1 \gamma_{x1} + \dots + m_{x-1} \gamma_{x, x-1} + m_x = 0 \\ m_1 \gamma_{x0} + m_2 \gamma_{x1} + \dots + m_x \gamma_{x, x-1} + m_{x+1} = 0 \\ \dots \dots \dots \\ m_{x-1} \gamma_{x0} + m_x \gamma_{x1} + \dots + m_{2x-2} \gamma_{x, x-1} + m_{2x-1} = 0 \end{cases}$$

Conversely, from (18) follows (16), hence the equations (18) must have exactly one solution  $\gamma_{x0}, \dots, \gamma_{x, x-1}$ .

Concerning the normalizing factor in  $\frac{1}{\lambda_x} \mathcal{Y}_x$ , we have

$$\begin{aligned} \lambda_x^2 &= \mathcal{Y}_x^2 = (\gamma_{x0} \mathcal{E}_0 + \gamma_{x1} \mathcal{E}_1 + \dots + \gamma_{x, x-1} \mathcal{E}_{x-1} + \mathcal{E}_x)^2 \\ &= (m_0 \gamma_{x0} + \dots + m_{x-1} \gamma_{x, x-1} + m_x) \gamma_{x0} \\ &\quad + (m_1 \gamma_{x0} + \dots + m_x \gamma_{x, x-1} + m_{x+1}) \gamma_{x1} \\ &\quad + \dots \\ &\quad + (m_x \gamma_{x0} + \dots + m_{2x-1} \gamma_{x, x-1} + m_{2x}) \gamma_{x, x-1} + m_{2x} \end{aligned}$$

and from (18):

$$\lambda_x^2 = m_x \gamma_{x0} + \dots + m_{2x-1} \gamma_{x, x-1} + m_{2x}.$$

With the abbreviations

$$X_0 = \mathcal{E}_0 \mathcal{E}, \dots, X_k = \mathcal{E}_k \mathcal{E}$$

we have

$$a_0 = \frac{1}{\lambda_0} X_0$$

$$a_1 = \frac{1}{\lambda_1} (\gamma_{10} X_0 + X_1)$$

$$\dots$$

$$a_k = \frac{1}{\lambda_k} (\gamma_{k0} X_0 + \gamma_{k1} X_1 + \dots + \gamma_{k, k-1} X_{k-1} + X_k).$$

For the calculation of  $a_0, a_1, \dots$  we recommend operating according to the following recipe, in the demonstration of which we confine ourselves to the most important case  $\kappa=3$ . The modulus procedendi for other values of the degree  $\kappa$  will be clear.

1. *Compute*

$$\xi_1 = \psi\left(\frac{1}{2n}\right), \quad \xi_2 = \psi\left(\frac{3}{2n}\right), \dots, \quad \xi_n = \psi\left(\frac{2n-1}{2n}\right).$$

In the interest of, the accuracy of the results it is advisable to take care that the equations

$$\frac{1}{n}(\xi_1 + \dots + \xi_n) = 0, \quad \frac{1}{n}(\xi_1^2 + \dots + \xi_n^2) = 1$$

are precisely or approximately satisfied. This will be the case if  $\mu_1=0, \mu_2=1$  hold; otherwise introduce

$$\xi_i^* = \gamma \xi_i + \beta, \dots, \quad \xi_n^* = \gamma \xi_n + \beta$$

instead of  $\xi_1, \dots, \xi_n$ , with convenient constants  $\gamma > 0, \beta$ .

2. *Compute*

$$m_0, m_1, \dots, m_6; \quad \chi_0, \dots, \chi_3; \quad \xi^2$$

Again it is useful to take care that the equations

$$\frac{1}{n}(\chi_1 + \dots + \chi_n) = 0; \quad \frac{1}{n}(\chi_1^2 + \dots + \chi_n^2) = 1$$

are precisely or approximately satisfied. This will be secured if

$\chi_1, \dots, \chi_n$  are distributed over nearly the same interval as

$$\xi_1, \dots, \xi_n.$$

3. *Form the scheme*

$$\begin{pmatrix} a_{00} & a_{01} & a_{02} & a_{03} \\ a_{10} & a_{11} & a_{12} & a_{13} \\ a_{20} & a_{21} & a_{22} & a_{23} \\ a_{30} & a_{31} & a_{32} & a_{33} \end{pmatrix} = \begin{pmatrix} m_0 & m_1 & m_2 & m_3 \\ m_1 & m_2 & m_3 & m_4 \\ m_2 & m_3 & m_4 & m_5 \\ m_3 & m_4 & m_5 & m_6 \end{pmatrix}$$

4a. To every element of the second, third and fourth row in this scheme add the corresponding element of the first row multiplied by  $-\frac{a_{10}}{a_{00}}, -\frac{a_{20}}{a_{00}}$ , and  $-\frac{a_{30}}{a_{00}}$  respectively, so that there results a scheme

$$\begin{pmatrix} a_{00} & a_{01} & a_{02} & a_{03} \\ 0 & a'_{11} & a'_{12} & a'_{13} \\ 0 & a'_{21} & a'_{22} & a'_{23} \\ 0 & a'_{31} & a'_{32} & a'_{33} \end{pmatrix}$$

4b. To every element of the second and third row in the scheme

$$\begin{pmatrix} a'_{11} & a'_{12} & a'_{13} \\ a'_{21} & a'_{22} & a'_{23} \\ a'_{31} & a'_{32} & a'_{33} \end{pmatrix}$$

add the corresponding element of the first row multiplied by  $-\frac{a'_{21}}{a'_{11}}$  and  $-\frac{a'_{31}}{a'_{11}}$  respectively, so that there results a scheme

$$\begin{pmatrix} a'_{11} & a'_{12} & a'_{13} \\ 0 & a''_{22} & a''_{23} \\ 0 & a''_{32} & a''_{33} \end{pmatrix}.$$

4c. To every element of the second row of the scheme

$$\begin{pmatrix} a''_{22} & a''_{23} \\ a''_{32} & a''_{33} \end{pmatrix}$$

add the corresponding element of the first row multiplied by  $-\frac{a''_{32}}{a''_{22}}$ , so that there results a scheme

$$\begin{pmatrix} a''_{22} & a''_{23} \\ 0 & a'''_{33} \end{pmatrix}.$$

5. Extract

$$\lambda_0 = \sqrt{a_{00}}, \quad \lambda_1 = \sqrt{a'_{11}}, \quad \lambda_2 = \sqrt{a''_{22}}, \quad \lambda_3 = \sqrt{a'''_{33}}.$$

6. Multiply the elements of the first, second and third row in the scheme

$$\begin{pmatrix} a_{00} & a_{01} & a_{02} & a_{03} \\ 0 & a'_{11} & a'_{12} & a'_{13} \\ 0 & 0 & a''_{22} & a''_{23} \\ 0 & 0 & 0 & a'''_{33} \end{pmatrix}$$

by  $\frac{1}{a_{00}}$ ,  $\frac{1}{a'_{11}}$ , and  $\frac{1}{a''_{22}}$  respectively, so that there results a scheme

$$B = \begin{pmatrix} 1 & b_{01} & b_{02} & b_{03} \\ 0 & 1 & b_{12} & b_{13} \\ 0 & 0 & 1 & b_{23} \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

and extract

$$\gamma_{10} = -b_{01}.$$

7. To every element of the first row in the scheme  $B$  add the corresponding element of the second row multiplied by  $-b_{01}$ , so

that there results a scheme

$$B' = \begin{pmatrix} 1 & 0 & \ell'_{02} & \ell'_{03} \\ 0 & 1 & \ell'_{12} & \ell'_{13} \\ 0 & 0 & 1 & \ell'_{23} \end{pmatrix},$$

and extract

$$\gamma_{20} = -\ell'_{02}, \quad \gamma_{21} = -\ell'_{12}.$$

8. To every element of the first and second row in the scheme  $B'$  add the corresponding element of the third row multiplied by  $-\ell'_{02}$  and  $-\ell'_{12}$  respectively, so that there results a scheme

$$B'' = \begin{pmatrix} 1 & 0 & 0 & \ell''_{03} \\ 0 & 1 & 0 & \ell''_{13} \\ 0 & 0 & 1 & \ell'_{23} \end{pmatrix},$$

and extract

$$\gamma_{30} = -\ell''_{03}, \quad \gamma_{31} = -\ell''_{13}, \quad \gamma_{32} = -\ell'_{23}.$$

9. Form

$$\begin{aligned} \gamma_0 &= X_0 \\ \gamma_1 &= \gamma_{10} X_0 + X_1 \\ \gamma_2 &= \gamma_{20} X_0 + \gamma_{21} X_1 + X_2 \\ \gamma_3 &= \gamma_{30} X_0 + \gamma_{31} X_1 + \gamma_{32} X_2 + X_3, \end{aligned}$$

$$M = a_0 = \frac{\gamma_0}{\lambda_0}, \quad \sigma = a_1 = \frac{\gamma_1}{\lambda_1}, \quad a_2 = \frac{\gamma_2}{\lambda_2}, \quad a_3 = \frac{\gamma_3}{\lambda_3}$$

$$S = \frac{a_2}{a_1}, \quad K = \frac{a_3}{a_1}$$

$$M_1 = \sqrt{\frac{a_1^2}{\ell^2 - a_0^2}}, \quad M_2 = \sqrt{\frac{a_1^2 + a_2^2}{\ell^2 - a_0^2}}, \quad M_3 = \sqrt{\frac{a_1^2 + a_2^2 + a_3^2}{\ell^2 - a_0^2}}$$

## 16. CONTROLS OF COMPUTATION

It is easy to point out controls for the process of evaluation of  $a_0, a_1, a_2, a_3$ , which do not require any considerable extra work, and yet indicate every occurring miscalculation with almost absolute safety. Such general controls, of course, can not bear upon the ascertainment of  $\xi_1, \dots, \xi_n$ .

A. Control of  $m_0, m_1, m_2, m_3$ ;

$$m_0 + 3(m_1 + m_2) + m_3 = \frac{1}{n} \sum_{\nu=1}^n (1 + \xi_\nu)^3.$$

B. Control of  $m_4, m_5, m_6$ ;

$$m_3 + 3(m_4 + m_5) + m_6 = \frac{1}{n} \sum_{\nu=1}^n \xi_\nu^3 (1 + \xi_\nu)^3.$$

C. Control of  $\tilde{X}_0, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3$ ;

$$\tilde{X}_0 + 3(\tilde{X}_1 + \tilde{X}_2) + \tilde{X}_3 = \frac{1}{n} \sum_{\nu=1}^n (1 + \xi_\nu)^3 \cdot x_\nu.$$

D. Control of  $\tilde{\xi}^2$ ;

$$1 + 2\tilde{X}_0 + \tilde{\xi}^2 = \frac{1}{n} \sum_{\nu=1}^n (1 + x_\nu)^2.$$

E. Control of  $\tilde{\gamma}_{10}, \tilde{\gamma}_{20}, \tilde{\gamma}_{21}; \tilde{\gamma}_{30}, \tilde{\gamma}_{31}, \tilde{\gamma}_{32}$ ;

The operations indicated under 3 - 8 in No. 15 are essentially nothing else than the solution of three systems of linear equations for one, two and three unknowns respectively, contracted into one uniform process of reckoning. Hence we can make use of the method of control by sums. We have to add the sums

$$S_0 = a_{00} + \dots + a_{03}$$

$$S_3 = a_{30} + \dots + a_{33}$$

as elements of a fifth column:

$$\begin{pmatrix} a_{00} & \dots & a_{03} & S_0 \\ \dots & \dots & \dots & \dots \\ a_{30} & \dots & a_{33} & S_3 \end{pmatrix}$$

and to transform this expanded scheme in the way described in No. 15. Then everywhere the sum of the first four elements of each row must equal its fifth element. If this is true for the scheme  $B''$  especially, it is practically impossible that  $\tilde{\gamma}_{10}, \dots, \tilde{\gamma}_{32}$  should have been wrongly computed.

F. Control of  $\tilde{\gamma}_0, \tilde{\gamma}_1, \tilde{\gamma}_2, \tilde{\gamma}_3$ ;

The computation of  $\tilde{\gamma}_0, \dots, \tilde{\gamma}_3$  should be performed by starting from the scheme

$$\begin{pmatrix} 1 & \gamma_{10} & \gamma_{20} & \gamma_{30} & S_0 \\ 0 & 1 & \gamma_{21} & \gamma_{31} & S_1 \\ 0 & 0 & 1 & \gamma_{32} & S_2 \\ 0 & 0 & 0 & 1 & S_3 \end{pmatrix}$$

with the meaning  $1 + \gamma_{10} + \gamma_{20} + \gamma_{30} = S_0$   
 $1 + \gamma_{21} + \gamma_{31} = S_1$   
 $1 + \gamma_{32} = S_2$   
 $1 = S_3.$

Multiply the elements of the first, second, third and fourth row by  $X_0, X_1, X_2$  and  $X_3$  respectively, and form the sums of the elements of each column. The sums of the first four columns are  $\gamma_0, \gamma_1, \gamma_2, \gamma_3$ , and we have the control  $\gamma_0 + \gamma_1 + \gamma_2 + \gamma_3 = R$ , where  $R$  designates the sum of the fifth column

## 17. EXAMPLES

I. Let the observations ( $n=30$ )

$\mathcal{C} = (-1.5, -1.0, -7, -5, -3, 0, +3, +6, +1.0, +1.8)$

be given, and let us first assume the normal type. The normal law of error being symmetric, we have  $m_1 = m_3 = m_5 = 0$ , and in this case we are able to write down the Tchebychef coefficients required:  $a_0 = X_0$ ,  $a_1 = \frac{X_1}{\sqrt{m_2}}$ ,  $a_2 = \frac{-m_2 X_0 + X_2}{\sqrt{m_4 - m_2^2}}$ ,  $a_3 = \frac{-m_4 X_1 + m_2 X_3}{\sqrt{m_6(m_2 m_4 - m_2^2)}}$ . Nevertheless we will proceed according to No. 15. But we will confine ourselves to give the resulting data of the different steps of computation only. A full reproduction of the complete process of reckoning is to be found in No. 19, dealing with a somewhat more general situation.

In the KELLEY-WOOD tables we find

$$\begin{aligned} \xi_1 &= -1.644854 & \xi_2 &= -1.036433 & \xi_3 &= -.764490 & \xi_4 &= -.385320 \\ \xi_5 &= -.125661 & \xi_6 &= .125661 & \xi_7 &= .385320 & \xi_8 &= .764490 \\ \xi_9 &= 1.036433 & \xi_{10} &= 1.644854. \end{aligned}$$



We obtain  $m_0 = 1$   $m_1 = 0$   $m_2 = +.87979$   $m_3 = 0$

$$m_4 = +1.74062 \quad m_5 = 0 \quad m_6 = +4.22829$$

$$X_0 = -.03000 \quad X_1 = +.87237 \quad X_2 = +.07317 \quad X_3 = +1.73577$$

$$\epsilon^2 = .87700;$$

$$\begin{pmatrix} a_{00} & a_{01} & a_{02} & a_{03} \\ 0 & a_{11} & a_{12} & a_{13} \\ 0 & 0 & a_{22} & a_{23} \\ 0 & 0 & 0 & a_{33} \end{pmatrix} = \begin{pmatrix} 1 & 0 & +.87779 & 0 \\ 0 & +.87979 & 0 & +1.74062 \\ 0 & 0 & +.96659 & 0 \\ 0 & 0 & 0 & +.78456 \end{pmatrix}$$

$$\lambda_0 = 1 \quad \lambda_1 = .93797 \quad \lambda_2 = .98315 \quad \lambda_3 = .88575$$

$$B = B' = \begin{pmatrix} 1 & 0 & +.87979 & 0 \\ 0 & 1 & 0 & +1.97845 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

$$\gamma_{10} = 0; \quad \gamma_{20} = -.87979 \quad \gamma_{21} = 0;$$

$$B'' = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & +1.97845 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

$$\gamma_{30} = 0 \quad \gamma_{31} = -1.97845 \quad \gamma_{32} = 0;$$

$$M = a_0 = -.03000 \quad \sigma = a_1 = +.93006 \quad a_2 = +.10127 \quad a_3 = +.01110$$

$$S = +.10889 \quad K = +.01193$$

$$M_1 = .99382 \quad M_2 = .99953 \quad M_3 = .99959$$

For comparison we give the value which is furnished by the traditional concept of dispersion:

$$\sqrt{\frac{1}{n} \sum (x_i - M)^2} = \sqrt{\epsilon^2 - a_0^2} = .93600.$$

II. Let the same observations as above be given, but now let us assume the step type. We can make use of the vectors in No. 13, which give at once

$$M = a_0 = -.03000 \quad \sigma = a_1 = +.92087 \quad a_2 = +.10184 \quad a_3 = +.12529$$

$$S = +.11059 \quad K = +.13606$$

$$M_1 = .98383 \quad M_2 = .98989 \quad M_3 = .99941$$

We note that for our observations  $\epsilon$  the normal type is, up to the degree 3, better than the step type.

## 18. ANALYSIS OF FREQUENCY GROUPS

In economic statistics, observations very often do not appear in the form dealt with in the preceding chapters. Instead, they usually are gathered into groups, so that there is given a set of values  $x_1, x_2, \dots, x_n$  and a set of corresponding positive values  $N_1, \dots, N_n$ , not necessarily integers. If  $N$  means the sum of  $N_1 + \dots + N_n$ , the ratios

$f_1 = \frac{N_1}{N}, \dots, f_2 = \frac{N_2}{N}, \dots, f_n = \frac{N_n}{N}$   
are called the "frequencies" of the "observations"  $x_1, \dots, x_n$ .  
The frequencies satisfy

$$f_1 > 0, f_2 > 0, \dots, f_n > 0 \quad \text{and} \quad f_1 + \dots + f_n = 1.$$

We shall now have to extend our developments to make them applicable in situations as stated above. To anyone who is familiar with integrals and sums in the sense of Stieltjes, it is clear that no special difficulty can arise.

Again we have to start from a frequency function  $\varphi(x)$ , and to agree which values  $\xi_1, \dots, \xi_n$  should be designated as "best values". Reflections similar to those of No. 2 make it reasonable to choose

$$\xi_1 = \psi\left(\frac{1}{2}f_1\right), \quad \xi_2 = \psi\left(f_1 + \frac{1}{2}f_2\right), \quad \xi_3 = \psi\left(f_1 + f_2 + \frac{1}{2}f_3\right), \\ \dots, \quad \xi_n = \psi\left(f_1 + f_2 + \dots + f_{n-1} + \frac{1}{2}f_n\right).$$

Apart from the best values, we only have to modify the definition of the product of vectors (No. 4). We define

$$\tilde{N} \cdot X = u_1 v_1 f_1 + u_2 v_2 f_2 + \dots + u_n v_n f_n.$$

If these modifications are kept in mind, all the definitions, theorems, proofs and remarks of Nos. 4 - 16 remain unaltered. Of course, the abbreviations  $m_\lambda$  and  $X_\lambda$  (No. 15) must now be read

$$m_\lambda = \xi_1^\lambda f_1 + \xi_2^\lambda f_2 + \dots + \xi_n^\lambda f_n \quad (\lambda = 0, 1, 2, \dots)$$

$X_\nu = \xi_1^\nu x_1 f_1 + \dots + \xi_n^\nu x_n f_n \quad (\nu = 0, 1, \dots, k),$   
and the controls A - D (No. 16):

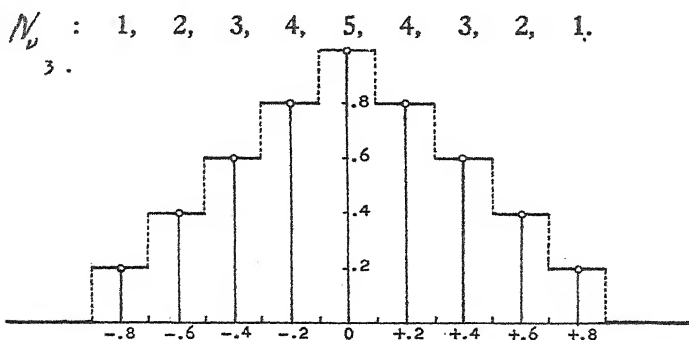
$$A. \quad r_0 + 3(m_1 + m_2) + m_3 = \sum_{\nu=1}^n (1 + \xi_\nu)^3 f_\nu$$

$$B. \quad m_3 + 3(m_4 + m_5) + m_6 = \sum_{\nu=1}^n (1 + \xi_{\nu})^3 \xi_{\nu}^3 f_{\nu}$$

$$C. \quad X_0 + 3(X_1 + X_2) + X_3 = \sum_{\nu=1}^n (1 + \xi_{\nu})^3 x_{\nu} f_{\nu}$$

$$D. \quad 1 + 2X_0 + X_0^2 = \sum_{\nu=1}^n (1 + x_{\nu})^2 f_{\nu}.$$

We are now in the position to illustrate the mechanism of skewness and kurtosis still more impressively than in No. 13. For this purpose we start from the frequency curve represented in Fig. 3; we choose  $n=9$  and



Then the best values become equidistant, and they are given by the abscissae of the points marked by small circles:

$$-.8, -.6, -.4, -.2, 0, +.2, +.4, +.6, +.8.$$

The ordinates  $n_{\nu}$  of these points are proportional to  $N_{\nu}$ , namely:

$$n_{\nu} = \frac{1}{0.2} f_{\nu} = 5 \frac{N_{\nu}}{N}.$$

The table below gives the corresponding vectors  $\beta_0, \dots, \beta_3$ , and also the vectors  $\beta_1 \pm \frac{1}{4} \beta_2$  and  $\beta_1 \pm \frac{1}{8} \beta_3$  as examples of distributions which show skewness or kurtosis in all purity.

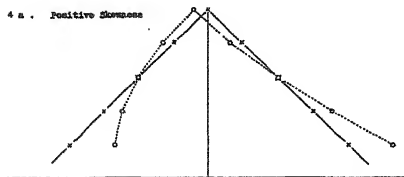
$\nu$	$\beta_0$	$\beta_1$	$\beta_2$	$\beta_3$	$\beta_1 + \frac{1}{4} \beta_2$	$\beta_1 - \frac{1}{4} \beta_2$	$\beta_1 + \frac{1}{8} \beta_3$	$\beta_1 - \frac{1}{8} \beta_3$
1	1	-2.0	+2.582	-2.561	-1.355	-2.645	-2.320	-1.680
2	1	-1.5	+1.076	+.116	-1.231	-1.769	-1.485	-1.515
3	1	-1.0	.000	+1.048	-1.000	-1.000	-.869	-1.131
4	1	-.5	-.645	+.815	-.661	-.339	-.398	-.602
5	1	.0	-.861	.000	-.215	+.215	.000	.000
6	1	+.5	-.645	-.815	+.339	+.661	+.398	+.602
7	1	+1.0	.000	-1.048	+1.000	+1.000	+.869	+1.131
8	1	+1.5	+1.076	-.116	+1.769	+1.231	+1.485	+1.515
9	1	+2.0	+2.582	+2.561	+2.645	+1.355	+2.320	+1.680

As in No. 13, let us illustrate the relations between the vector  $\bar{z}_1$  and the vectors  $\bar{z}_1 \pm \frac{1}{4} \bar{z}_2$  and  $\bar{z}_1 \pm \frac{1}{8} \bar{z}_3$  by means of some figures. This time however, we shall not only consider the components of the vectors, but also operate with the values  $f_\nu$ . We do that by associating every vector  $(u_1, \dots, u_n)$  with the system of points

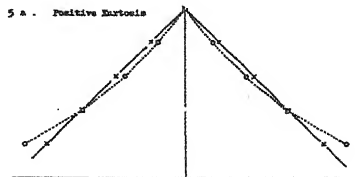
$$(u_1, f_1), (u_2, f_2), \dots, (u_n, f_n).$$

Thus, in the figures 4a - 5b, the vector  $\bar{z}_1$  is every time associated with the system of points marked by crosses, whereas the system of points marked by circles successively correspond to the vectors  $\bar{z}_1 \pm \frac{1}{4} \bar{z}_2$  and  $\bar{z}_1 \pm \frac{1}{8} \bar{z}_3$ .

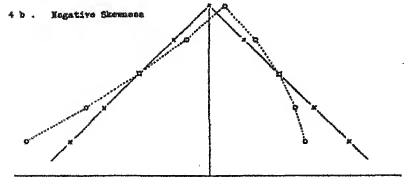
4 a . Positive Skewness



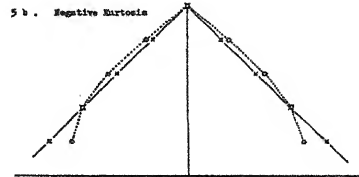
5 a . Positive Kurtosis



4 b . Negative Skewness



5 b . Negative Kurtosis



The statements in No. 13 concerning the meaning of skewness as overconcentration to the left or to the right, and of kurtosis as overconcentration or underconcentration near the mean should be recognized.

Until now, the values  $N_\nu$  were supposed to be really positive,

but there is no difficulty in allowing some of them to equal zero. Then, it is true, the formulation of some intermediary theorems must be changed. Yet, the existence and the main properties of the Tchebycheff coefficients remain untouched, and *their values are independent of those  $x_\nu$  for which the corresponding  $f_\nu$  are equal to zero*. To know this is sometimes useful in order to get a scheme of computation of the highest possible uniformity.

### 19. EXAMPLE

To conclude, we reproduce the reckoning of an example, frequently discussed, concerning observations of the right ascension of the pole star (see: A. L. BOWLEY, *Elements of Statistics*, 4th ed., p. 255). The given data are

$x_\nu^*$ :	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6
$N_\nu$ :	1	6	12	21	36	61	73	82	72	63	38	16	5	1

and the normal type shall be assumed.

Because the function

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

satisfies  $\mu_1 = 0, \mu_2 = 1$ , it will be suitable to start from the best values of this specimen. These best values  $\xi_1, \dots, \xi_{14}$  stretch from -3 to +3 approximately. In order to have the values  $x_\nu$ , with which we intend to work, in coextension with  $\xi_1, \dots, \xi_{14}$ , we choose

$$x_\nu = \frac{1}{2} x_\nu^* \quad \text{i.e.} \quad x_\nu^* = 2 x_\nu \quad (\nu = 1, \dots, 14).$$

Between the means  $M$ ,  $M^*$  and the dispersions  $\sigma$ ,  $\sigma^*$  of the observations  $x_\nu$ ,  $x_\nu^*$  there exist the connections (theorem 9)

$$M^* = 2M, \quad \sigma^* = 2\sigma,$$

whereas the measures of skewness and kurtosis as well as the measures of approximation do not change at the transition from  $x_\nu$  to  $x_\nu^*$  (theorems 11 and 12).

Computation of  $m_0, \dots, m_6$ ;  $\chi_0, \dots, \chi_3$ ;  $\epsilon^z$ .

$\nu$	$\xi_\nu$	$f_\nu$	$\xi_\nu f_\nu$	$\xi_\nu^2 f_\nu$	$\xi_\nu^3 f_\nu$
1	-3.08242	.00205 339	-.00632 941	.01950 990	-.06013 77
2	-2.39928 8	.01232 033	-.02956 002	.07092 300	-.17016 47
3	-1.93176 9	.02464 066	-.04760 006	.09195 232	-.17763 06
4	-1.54996 6	.04312 11	-.06683 62	.10359 38	-.16056 69
5	-1.17951 6	.07392 20	-.08719 22	.10284 46	-.12130 67
6	-.77663 9	.12525 7	-.09727 9	.07555 1	-.05867 6
7	-.36846 6	.14989 7	-.05523 2	.02035 1	-.00749 9
8	+.03861 3	.16837 8	+.00650 2	.00025 1	+.00001 0
9	+.44963 0	.14784 4	+.06647 5	.02988 9	+.01343 9
10	+.88571 7	.12936 34	+.11457 94	.10148 49	+.08988 69
11	+1.37743 5	.07802 87	+.10747 95	.14804 60	+.20392 37
12	+1.89953 0	.03285 421	+.06240 756	.11854 503	+.22517 98
13	+2.44778 6	.01026 694	+.02513 127	.06151 597	+.15057 79
14	+3.08242	.00205 339	+.00632 941	.01950 990	+.06013 77
		+1.00000 = $m_0$	-.00113 = $m_1$	+.96397 = $m_2$	-.01283 = $m_3$

$\nu$	$\xi_\nu^4 f_\nu$	$\xi_\nu^5 f_\nu$	$\xi_\nu^6 f_\nu$
1	.18536 96	-.57138 7	1.76125 5
2	.40827 31	-.97956 7	2.35026 3
3	.34314 13	-.66287 0	1.28051 2
4	.24887 3	-.38574 5	.59789 2
5	.14308 3	-.16876 9	.19906 6
6	.04537 0	-.03539 1	.02748 6
7	.00276 3	-.00101 8	.00037 5
8	.00000 0	+.00000 0	.00000 0
9	.00604 3	+.00271 7	.00122 2
10	.07961 4	+.07051 5	.06245 6
11	.28089 2	+.38691 0	.53294 3
12	.42773 58	+.81249 7	1.54336 2
13	.36858 25	+.90221 1	2.20841 9
14	.18536 96	+.57138 7	1.76125 5
		+ 2.72531 = $m_4$	- .05851 = $m_5$
			+ 12.32651 = $m_6$

$\nu$	$X_\nu$	$X_\nu f_\nu$	$\xi_\nu X_\nu f_\nu$	$\xi_\nu^2 X_\nu f_\nu$	$\xi_\nu^3 X_\nu f_\nu$	$X_\nu^2 f_\nu$
1	-3.5	-.00718 687	.02215 295	-.06828 47	.21048 2	.02515 4
2	-3.0	-.03696 099	.08868 006	-.21276 90	.51049 4	.11088 3
3	-2.5	-.06160 164	.11900 014	-.22988 08	.44407 7	.15400 4
4	-2.0	-.08624 23	.13367 26	-.20718 80	.32113 4	.17248 5
5	-1.5	-.11088 30	.13078 83	-.15426 69	.18196 0	.16632 4
6	-1.0	-.12525 67	.09727 9	-.07555 1	.05867 6	.12525 7
7	-.5	-.07494 9	.02761 6	-.01017 6	.00374 9	.03747 4
8	.0	.00000 0	.00000 0	.00000 0	.00000 0	.00000 0
9	.5	+.07392 2	.03323 8	+.01494 5	.00672 0	.03696 1
10	+1.0	+.12936 34	.11457 94	+.10148 49	.08988 7	.12936 3
11	+1.5	+.11704 31	.16121 93	+.22206 91	.30588 6	.17556 5
12	+2.0	+.06570 842	.12481 512	+.23709 01	.45036 0	.13141 7
13	+2.5	+.02566 735	.06282 818	+.15378 99	.37644 5	.06416 8
14	+3.0	+.00616 016	.01898 820	+.05852 96	.18041 3	.01848 0
		-.08522 = $X_0$	+1.13486 = $X_1$	-.17021 = $X_2$	+3.14028 = $X_3$	1.34754 = $\xi^2$

## Controls A - D.

$\nu$	$1 + \xi_\nu$	$(1 + \xi_\nu)^3$	$(1 + \xi_\nu)^3 f_\nu$	$(1 + \xi_\nu)^3 \xi_\nu^3 f_\nu$	$(1 + \xi_\nu)^3 X_\nu f_\nu$	$(1 + X_\nu)^2 f_\nu$
1	-2.08242	-9.0304	-.01854 3	+.54306 5	+.06490 0	.01283 4
2	-1.39928 8	-2.73982	-.03375 5	+.46622 0	+.10126 6	.04928 1
3	-.93176 9	-.80896	-.01993 3	+.14369 5	+.04983 3	.05544 1
4	-.54996 6	-.16634	-.00717 3	+.02670 9	+.01434 6	.04312 1
5	-.17951 6	-.00579	-.00042 8	+.00070 2	+.00064 2	.01848 1
6	+.22336 1	+.01114	+.00139 6	-.00065 4	-.00139 5	.00000 0
7	+.63153 4	+.25188	+.03775 6	-.00188 9	-.01887 8	.03747 4
8	+1.03861 3	+1.12037	+.18864 5	+.00001 1	+.00000 0	.16837 8
9	+1.44963 0	+3.04629	+.45037 6	+.04093 9	+.22518 8	.33264 9
10	+1.88571 7	+6.70548	+.86744 3	+.60273 4	+.86744 4	.51745 4
11	+2.37743 5	+13.43773	+1.04852 9	+2.74027 2	+1.57279 4	.48767 9
12	+2.89953 0	+24.37714	+.80089 2	+5.48924 0	+1.60178 3	.29568 8
13	+3.44778 6	+40.98462	+.42078 7	+6.17137 8	+1.05196 7	.12577 0
14	+4.08242	+68.0382	+.13970 9	+4.09166 2	+.41912 6	.03285 4
			+3.87570	+20.31408	+5.94902	2.17710

$$m_1 + 3(m_1 + m_2) + m_3 = 3.87569$$

$$m_2 + 3(m_4 + m_5) + m_6 = 20.31408$$

$$X_0 + 3(X_1 + X_2) + X_3 = 5.94901$$

$$1 + 2X_0 + \xi^2 = 2.17710$$

Computation of  $a_{00}, \dots, a_{03}; a'_{11}, \dots, a'''_{33}$  (twice underlined)  
and of  $\lambda_0, \dots, \lambda_3$ , with control by sums.

1	- .00113	+ .96397	- .01283	+ 1.95001
- .00113	+ .96397	- .01283	+ 2.72531	+ 3.67532
	- .00000	+ .00109	- .00001	+ .00220
	+ .96397	- .01174	+ 2.72530	+ 3.67752
+ .96397	- .01283	+ 2.72531	- .05851	+ 3.61794
	+ .00109	- .92924	+ .01237	- 1.87975
	- .01174	+ 1.79607	- .04614	+ 1.73819
- .01283	+ 2.72531	- .05851	+ 12.32651	+ 14.98048
	- .00001	+ .01237	- .00016	+ .02502
	+ 2.72530	- .04614	+ 12.32635	+ 15.00550
	+ .96397	- .01174	+ 2.72530	+ 3.67752
	- .01174	+ 1.79607	- .04614	+ 1.73819
		- .00014	+ .03319	+ .04479
		+ 1.79593	- .01295	+ 1.78298
+ 2.72530	- .04614	+ 12.32635	+ 15.00550	
	+ .03319	- 7.70486	- 10.39694	
	- .01295	+ 4.62149	+ 4.60856	
	+ 1.79593	- .01295	+ 1.78298	
	- .01295	+ 4.62149	+ 4.60856	
		- .00009	+ .01286	
		+ 4.62140	+ 4.62142	

$$\lambda_0 = 1$$

$$\lambda_1 = .98182$$

$$\lambda_2 = 1.34012$$

$$\lambda_3 = 2.14974$$



Computation of  $\gamma_{10}, \dots, \gamma_{32}$ , with control by sums.

$$\gamma_{10} = +.00113$$

1	-.00113	+.96397	-.01283	+1.95001
	+.00113	-.00001	+.00319	+.00431
1	0	+.96396	-.00964	+1.95432
<hr/>				
1	-.01218	+2.82716		+2.81497
<hr/>				
	1	-.00721		+.99279

$$\gamma_{20} = -.96396$$

$$\gamma_{21} = +.01218$$

1	0	+.96396	-.00964	+1.95432
		-.96396	+.00695	-.95701
1	0	0	-.00269	+.99731
<hr/>				
1	-.01218	+2.82716		+3.81497
	+.01218	-.00009		+.01209
1	0	+2.82707		+3.82706
<hr/>				
	1	-.00721		+.99279

$$\gamma_{30} = +.00269$$

$$\gamma_{31} = -2.82707$$

$$\gamma_{32} = +.00721$$

Computation of  $\gamma_0, \dots, \gamma_3$ , with control by sums.

1	+.00113	-.96396	+.00269	+ .03986
	1	+.01218	-2.82707	-1.81489
		1	+.00721	+1.00721
			1	+1.00000
<hr/>				
-.08522	-.00010	+.08214	-.00023	+ .00340
	+1.13486	+.01382	-3.20833	-2.05965
		-.17021	-.00123	-.17144
			+3.14028	+3.14028
<hr/>				
-.08522	+1.13476	-.07425	-.06951	+ .90579
= $\gamma_0$	= $\gamma_1$	= $\gamma_2$	= $\gamma_3$	
<hr/>				
$\gamma_0 + \dots + \gamma_3 = +.90578$				

Finishing computations.

$$\begin{array}{llll} M = a_0 = .08522 & \sigma = a_1 = 1.15577 & a_2 = -.05541 & a_3 = -.03234 \\ M^* = .17044 & \sigma^* = 2.31154 & S = -.04794 & K = -.02799 \end{array}$$

$$\begin{array}{llll} \epsilon^2 = 1.34754 & a_1^2 = 1.33580 & M_1^2 = .99666 & M_1 = .99833 \\ a_0^2 = .00726 & a_2^2 = .00307 & & \\ \epsilon^2 - a_0^2 = 1.34028 & a_1^2 + a_2^2 = 1.33887 & M_2^2 = .99895 & M_2 = .99947 \\ & a_3^2 = .00105 & & \\ & a_1^2 + a_2^2 + a_3^2 = 1.33992 & M_3^2 = .99973 & M_3 = .99987 \end{array}$$

So long as we pay regard to the Tchebycheff coefficients  $a_0, \dots, a_3$  only, the purport of our results is that *the observations are somewhat overconcentrated to the right, and somewhat underconcentrated near the mean.* The sum of the squares of the Tchebycheff coefficients with higher subscripts than 3 is

$$a_4^2 + \dots + a_{13}^2 = \epsilon^2 - (a_0^2 + \dots + a_3^2) = .00036;$$

it is small compared with  $a_2^2 = .00307$  and  $a_3^2 = .00105$ . The vectors  $j_0, \dots, j_{13}$  being normalized, we are sure that *the influence of  $a_4, \dots, a_{13}$  cannot essentially disturb our statements.*

Finally we give an illustration by computing and drawing the "best curve" of the normal type, corresponding to the observations  $\chi_\nu$ . With it we mean that curve  $y = \frac{1}{\sigma} \varphi\left(\frac{x-\beta}{\sigma}\right)$ , the best values of which are the components of the vector  $a_0 j_0 + a_1 j_1$ . The values  $\gamma, \beta$  (see No. 2) have to satisfy

$$\beta \epsilon_0 + \gamma \epsilon_1 = a_0 j_0 + a_1 j_1;$$

substituting

$$j_0 = \frac{1}{\lambda_0} \epsilon_0, \quad j_1 = \frac{1}{\lambda_1} (\gamma_{10} \epsilon_0 + \epsilon_1)$$

we get

$$\left\{ \beta - \left( \frac{a_0}{\lambda_0} + \gamma_{10} \frac{a_1}{\lambda_1} \right) \right\} \epsilon_0 + \left\{ \gamma - \frac{a_1}{\lambda_1} \right\} \epsilon_1 = 0,$$

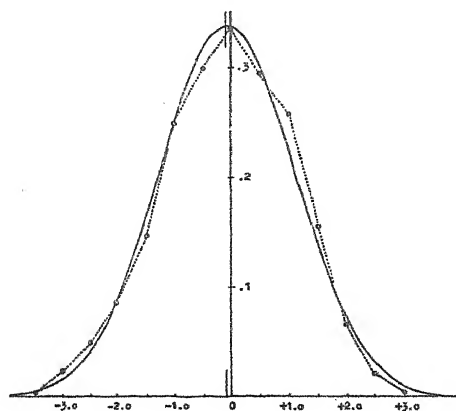
hence

$$\gamma = \frac{a_1}{\lambda_1} = +1.17717$$

$$\beta = \frac{a_0}{\lambda_0} + \gamma \frac{a_1}{\lambda_1} = - .08389.$$

With these values  $\gamma$  and  $\beta$ , the curve in Fig. 6 represents the function

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\beta)^2}{2\sigma^2}}$$



The abscissae of the points marked by circles are the observations  $x_v$ , their ordinates are equal to the corresponding  $f_v$  divided by the length 0.5 of the group intervals.

University of Kiel, Germany.

## EDITORIAL

---

### A New Type of Average for Security Prices

The market averages that are most popular with the American investing public are essentially weighted or unweighted means of security prices at designated intervals. As a rule, they ignore the volume of sales—an element to which experienced traders attribute considerable importance. Such averages endeavor only to reflect the average price level at periodic intervals, and all of those published are entirely satisfactory in this respect.

In this note we shall discuss an *acquisition average* which, instead of being concerned with the price level at a given moment attempts to answer the question, "what is the average price actually paid for the securities by their present owners."

The problem can best be appreciated by presenting two examples of acquisition averages prior to the mathematical theory. The first entry of Table 1 states that for the week ending January 7, 1928 United States Steel common closed at 150 6-8, and that the acquisition average on this date was \$137.75. At the time of the market crash in October, 1929, the acquisition average had risen to about \$212, and at the present moment this average has receded to about \$48. Of course, some of the individuals who bought Steel at about \$200 per share are still holding on to it, whereas others among the present holders obtained theirs in 1932 at less than \$25 per share. According to our theory, the mean of such acquisition prices is the \$48 noted above.

As an illustration of corresponding averages computed on a *daily* basis, table 2 presents the daily closing prices and acquisition averages for Auburn, covering the last half of 1934. This stock was selected because of its relatively small capitalization and frequent activity.

TABLE 1.

WEEKLY CLOSING PRICES AND ACQUISITION AVERAGES FOR U. S. STEEL,  
1928-1933.

1928	Close	A.A.	1929	Close	A.A.	1930	Close	A.A.
1- 7	150.6	137.75	1- 5	161.6	152.34	1- 4	169.1	192.63
14	146.0	138.2	12	165.5	133.09	11	169.5	191.32
21	147.4	138.80	19	188.4	156.44	18	169.6	190.51
28	146.3	139.19	26	187.1	160.23	25	179.0	189.43
2- 4	143.4	139.42	2- 2	184.7	162.07	2- 1	184.4	188.86
11	145.7	139.69	9	173.4	163.17	8	182.6	188.59
18	140.0	139.70	16	169.3	163.64	15	186.0	188.37
25	140.0	139.71	23	182.0	164.51	22	183.0	188.15
3- 3	140.2 x	138.06	3- 3	188.4 x	165.62	3- 1	184.3 x	186.21
10	144.7	138.33	9	185.2	166.98	8	181.0	186.07
17	147.7	138.99	16	188.0	168.58	15	179.2	185.88
24	147.5	139.83	23	181.1	169.51	22	187.6	185.68
31	147.3	140.35	30	183.6	170.46	29	193.4	185.97
4- 7	147.1	140.51	4- 6	186.3	172.06	4- 5	196.4	186.50
14	150.0	141.09	13	188.5	173.52	12	193.1	186.98
21	145.5	141.34	20	186.0	174.41	19	195.2	187.30
28	145.3	141.48	27	186.2	174.79	26	188.0	187.51
5- 5	148.0	141.58	5- 4	182.1	175.20	5- 3	170.2	186.73
12	148.6	141.90	11	179.4	175.33	10	172.6	185.65
19	145.6	142.10	18	174.6	175.38	17	172.7	185.30
26	146.7	142.22	25	167.7	175.25	24	172.0	184.80
6- 2	145.2 x	140.70	6- 1	165.0 x	173.28	31	173.5 x	182.87
9	140.6	140.82	8	168.0	173.16	6- 7	164.2	182.48
16	138.5	140.75	15	175.5 x	169.40	14	162.4	181.42
23	133.2	140.52	22	180.6	169.80	21	155.2	179.79
30	137.3	140.39	29	190.6	170.90	28	156.0	178.52
7- 7	140.0	140.35	7- 6	196.3	172.43	7- 5	157.7	177.99
14	135.0	140.26	13	202.3	174.63	12	160.6	177.46
21	138.2	140.16	20	207.7	177.48	19	166.6	177.03
28	144.2	140.21	27	206.0	179.85	26	169.7	176.83
8- 4	140.3	140.28	8- 3	204.4	182.20	8- 2	166.2	176.61
11	142.7	140.32	10	218.0	185.68	9	159.4	176.03
18	149.0	140.78	17	238.5	192.17	16	165.3	175.52
25	151.2	141.45	24	258.2	197.54	23	168.2	175.29
9- 1	154.0 x	140.51	31	256.4 x	198.64	30	171.2 x	173.38
8	155.7	141.54	9- 7	247.4	202.19	9- 6	173.1	173.34
15	159.0	142.71	14	233.2	205.43	13	170.2	173.28
22	157.6	143.62	21	232.1	207.70	20	163.7	173.00
29	159.2	145.12	28	225.0	210.12	27	158.2	172.21
10- 6	158.3	146.31	10- 5	217.6	211.40	10- 4	156.6	171.31
13	164.6	147.56	12	230.6	212.56	11	148.4	169.46
20	162.0	148.26	19	209.0	213.41	18	145.3	168.25
27	161.5	149.16	26	203.4	212.38	25	151.4	167.23
11- 3	160.7	149.68	11- 2	193.2	211.06	11- 1	145.6	166.39
10	164.3	150.34	9	171.0	209.67	8	140.4	165.39
17	171.2	152.26	16	164.2	207.09	15	147.7	163.84
24	167.6	152.95	23	167.0	205.42	22	147.2	163.12
12- 1	165.0 x	151.50	30	162.1 x	201.81	29	145.4	162.59
8	151.2	151.95	12- 7	182.6	200.30	12- 6	142.3 x	160.38
15	151.1	151.92	14	174.0	196.87	13	136.4	159.24
22	156.0	151.97	21	163.0	195.11	20	140.5	158.19
29	159.4	152.11	28	164.4	193.76	27	136.6	157.73

x = ex-dividend.

x = rights.

TABLE 1—(Continued)

1931	Close	A.A.	1932	Close	A.A.	1933	Close	A.A.
1- 3	143.4	157.38	1- 2	37.1	83.57	1- 7	29.7	46.40
10	143.7	157.06	9	42.7	80.59	14	29.6	46.10
17	139.1	156.65	16	44.1	78.50	21	28.5	45.90
24	142.3	156.30	23	41.4	77.00	28	27.7	45.74
31	139.2	155.99	30	37.3	75.28	2- 4	26.3	45.54
2- 7	140.3	155.67	2- 6	38.5	73.93	11	28.3	45.36
14	145.2	155.30	13	49.0	72.23	18	26.7	45.20
21	148.6	155.06	20	48.4	70.84	25	24.5	44.97
28	147.4	154.87	27	47.0	70.20	3- 4	26.2	44.68
3- 7	146.6 x	152.95	3- 5	50.6 x	69.00	11		
14	144.4	152.82	12	46.7	68.44	18	30.6	44.33
21	147.4	152.65	19	41.5	67.33	25	28.5	44.13
28	141.4	152.50	26	40.2	66.50	4- 1	27.5	44.00
4- 4	140.0	152.23	4- 2	39.0	65.47	8	30.3	43.82
11	137.3	151.80	9	34.6	64.16	15	32.4	43.57
18	132.6	151.22	16	33.2	63.15	22	42.3	43.23
25	125.5	150.14	23	29.1	62.33	29	46.5	43.30
5- 2	115.2	148.10	30	28.2	61.60	5- 6	46.7	43.48
9	111.5	146.24	5- 7	30.0	60.80	13	47.4	43.66
16	101.5	143.97	14	27.0	60.29	20	47.2	43.78
23	98.4	141.33	21	29.0	59.81	27	53.0	44.05
30	91.0 x	136.40	28	27.2	59.34	6- 3	52.1	44.29
6- 6	89.3	132.63	6- 4	30.2	58.58	10	55.4	44.58
13	90.7	130.23	11	26.5	57.57	17	53.1	44.97
20	92.7	128.77	18	25.4	57.04	24	57.4	45.34
27	104.3	125.91	25	23.4	56.62	7- 1	59.7	45.84
7- 4	105.0	124.96	7- 2	23.6	56.13	8	65.2	46.32
11	96.4	123.23	9	21.6	55.83	15	64.2	46.88
18	94.4	121.56	16	23.3	55.43	22	52.2	47.39
25	90.2	120.50	23	24.7	55.18	29	54.3	47.49
8- 1	85.7	118.79	30	28.7	54.51	8- 5	51.4	47.56
8	86.0	118.04	8- 6	41.4	53.70	12	53.4	47.60
15	93.0	116.89	13	37.4	52.86	19	52.7	47.66
22	87.4	116.23	20	40.7	52.38	26	58.4	47.79
29	90.5	115.66	27	48.3	51.91	9- 2	55.3	47.91
9- 5	83.0 x	113.91	9- 3	51.4	51.81	9	51.5	47.96
12	80.5	112.74	10	48.6	51.73	16	55.0	48.05
19	75.2	110.78	17	38.7	51.24	23	49.4	48.17
26	77.1	109.00	24	45.4	50.82	30	45.5	48.16
10- 3	68.4	107.16	10- 1	43.7	50.63	10- 7	47.3	48.14
10	70.7	104.91	8	35.4	50.18	14	43.2	48.11
17	68.6	103.87	15	37.7	49.76	21	35.2	47.82
24	71.5	102.97	22	35.1	49.39	28	39.2	47.57
31	67.4	101.79	29	35.3	49.12	11- 4	40.5	47.43
11- 7	72.3	100.88	11- 5	35.0	48.87	11	42.2	47.35
14	67.7	99.85	12	39.2	48.53	18	43.3	47.26
21	60.6	98.59	19	36.1	48.32	25	45.0	47.21
28	53.5	96.92	26	32.7	48.16	12- 2	44.6	47.18
12- 5	54.0 x	93.36	12- 3	30.4	47.90	9	47.4	47.17
12	44.0	90.78	10	32.2	47.61	16	45.6	47.16
19	41.4	87.09	17	30.1	47.33	23	47.7	47.15
26	37.6	85.55	24	26.5	46.95	30	47.6	47.16
			31	27.4	46.63			

x = ex-dividend.

TABLE 2.

DAILY CLOSING PRICES AND ACQUISITION AVERAGES FOR AUBURN,  
JULY 1ST—DECEMBER 30TH, 1933.

1933 July Aug.	Close	A.A.	1933 Sept. Oct.	Close	A.A.	1933 Nov. Dec.	Close	A.A.
1	66.0	61.59	1	61.4	62.13	1	36.4	54.57
3	69.7	62.00	5	58.4	62.12	2	37.0	54.55
5	68.4	62.43	6	59.4	62.06	3	38.6	54.48
6	68.6	62.59	7	59.0	62.04	4	39.0	54.39
7	67.2	62.87	8	58.2	61.99	6	39.0	54.34
8	67.2	62.91	9	58.3	61.99	8	43.0	53.74
10	67.4	63.06	11	62.2	61.94	9	43.0	53.46
11	68.0	63.38	12	61.0	61.93	10	41.0	53.40
12	78.6	66.47	13	61.4	61.93	11	42.4	53.32
13	77.3	68.70	14	61.6	61.92	13	43.0	53.24
14	75.0	69.23	15	59.7	61.90	14	38.0	53.06
15	76.5	69.40	16	62.0	61.90	15	36.0	52.96
17	80.0	70.36	18	59.5	61.88	16	43.6	52.78
18	78.0	70.73	19	60.0	61.83	17	43.0	52.66
19	70.6	71.06	20	56.6 x	61.25	18	43.0	52.63
20	58.1	70.13	21	50.6	60.93	20	46.6	52.45
21	50.0	68.48	22	52.4	60.71	21	45.1	52.31
22	46.4	67.38	23	51.4	60.61	22	45.0	52.26
24	54.4	66.61	25	49.4	60.44	23	43.2	52.13
25	52.5	65.99	26	47.6	60.21	24	45.0	52.07
26	54.2	65.72	27	46.5	59.90	25	44.4	52.07
27	58.0	65.33	28	47.2	59.81	27	42.2	51.98
28	55.4	65.22	29	46.5	59.68	28	43.0	51.91
31	52.1	64.57	30	46.0	59.55	29	44.6	51.89
1	54.6	64.37	2	46.0	59.46	1	45.2	51.81
2	57.2	64.09	3	45.0	59.35	2	44.7	51.80
3	54.6	63.95	4	50.4	59.32	4	45.1	51.77
4	53.4	63.85	5	48.0	59.22	5	48.0	51.68
7	53.4	63.81	6	48.0	59.19	6	46.4	51.64
8	56.0	63.57	7	48.4	59.14	7	48.6	51.46
9	61.4	63.32	9	49.4	59.08	8	49.6	51.35
10	58.4	63.16	10	49.0	59.03	9	56.4	51.51
11	57.0	63.10	11	48.4	58.94	11	57.0	51.95
14	57.6	63.01	13	46.2	58.89	12	55.4	52.10
15	59.0	62.98	14	45.6	58.81	13	54.4	52.17
16	55.0	62.84	16	43.0	58.67	14	57.0	52.41
17	60.6	62.72	17	41.6	58.33	15	57.2	52.76
18	57.4	62.65	18	38.0	57.96	16	55.7	52.89
21	59.0	62.60	19	35.0	56.87	18	55.0	53.04
22	61.0	62.52	20	37.2	56.38	19	53.5	53.08
23	59.0	62.44	21	34.0	56.05	20	49.4 x	52.48
24	58.6	62.38	23	36.0	55.45	21	49.4	52.42
25	61.4	62.25	24	38.0	55.25	22	54.4	52.40
28	62.0	62.23	25	38.6	54.96	23	53.3	52.42
29	61.0	62.19	26	37.0	54.91	26	52.0	52.42
30	60.2	62.16	27	37.4	54.86	27	52.4	52.42
31	59.4	62.14	28	36.7	54.82	28	54.0	52.43
			30	35.4	54.73	29	54.0	52.45
			31	35.0	54.63	30	54.6	52.50

x = ex-dividend.

We shall now develop the theory on which the preceding tables were constructed. As a simple illustration let us suppose that 100 individuals start an enterprise, that a total of 100 shares of stock are issued, and that each of the individuals purchases one share for \$100. The total book value of the issue at the date of issue is therefore,

$$V_0 = \$100 \times 100 = \$10\,000,$$

and the acquisition average then is

$$A_0 = \frac{V_0}{100} = \$100.00.$$

If the first transfer of stock resulted from the sale of a single share at 150, the total amount paid by the group now owning all the issue is obviously

$$V_1 = 99(100) + 150 = 10\,050,$$

and the new acquisition average is

$$A_1 = \frac{V_1}{100} = (1 - \frac{1}{100})A_0 + \frac{p_1}{100} = 100.50.$$

If somewhat later the next sale of stock is a single share at 50, we may assume that

$$V_2 = 99(100.50) + 50 = 9999.50$$

and consequently

$$A_2 = \frac{V_2}{100} = (1 - \frac{1}{100})A_1 + \frac{p_2}{100} = 99.995.$$

Our first assumption is, therefore, that *whenever the sale of a share of stock is recorded, it is equally likely that any one of the previous holders sold the share*. More will be said of this assumption later.

In generalizing, let us adopt the following notation:

$L$  designates the number of share units listed for an issue

$A_0$  is the acquisition average at a given initial date.

$p_x$  denotes the price at which the  $x$ -th unit of stock is sold following the initial date.



$A_x$  is the acquisition average immediately after the sale of the  $x$ -th unit.

We have then that

$$\begin{aligned} A_1 &= (1 - \frac{1}{L}) A_0 + \frac{p_1}{L} \\ A_2 &= (1 - \frac{1}{L}) A_1 + \frac{p_2}{L} = (1 - \frac{1}{L})^2 A_0 + \frac{p_1}{L} (1 - \frac{1}{L}) + \frac{p_2}{L} \\ A_3 &= (1 - \frac{1}{L}) A_2 + \frac{p_3}{L} = (1 - \frac{1}{L})^3 A_0 + \frac{p_1}{L} (1 - \frac{1}{L})^2 + \frac{p_2}{L} (1 - \frac{1}{L}) + \frac{p_3}{L} \\ &\dots \\ (1) \quad A_x &= (1 - \frac{1}{L})^x A_0 + \frac{p_1}{L} (1 - \frac{1}{L})^{x-1} + \frac{p_2}{L} (1 - \frac{1}{L})^{x-2} + \dots + \frac{p_{x-1}}{L} (1 - \frac{1}{L}) + \frac{p_x}{L} \end{aligned}$$

If we multiply both sides of this last equation by  $(1 - \frac{1}{L})$  and then subtract the resulting equation from (1), we obtain

$$(2) \quad A_x = A_0 (1 - \frac{1}{L})^x - p_0 (1 - \frac{1}{L})^x + (1 - \frac{1}{L})^{x-1} (p_1 - p_0) + (1 - \frac{1}{L})^{x-2} (p_2 - p_1) + \dots + (1 - \frac{1}{L}) (p_{x-1} - p_{x-2}) + p_x.$$

We shall now make a second assumption, namely, that the prices vary linearly from 0 to  $x$ . To illustrate, if Steel closes one week at 54 and during the next week 100,000 shares are sold after which the close is 59, our assumption means that after 20,000 shares were sold the quotation is 55, at 40,000 shares the price is 56, etc. Actually the price trend between two dates is not a straight line but rather a scattering of points. However, the linear assumption introduces compensating errors which have been found to result in only negligible variations in the resulting acquisition averages. We may write, therefore,

$$(3) \quad \begin{cases} p_i = p_0 + \frac{p_x - p_0}{x} i \\ p_{i+1} - p_i = \frac{p_x - p_0}{x} \end{cases}$$

and equation (2) then reduces to

$$(4) \quad A_x = A_0 (1 - \frac{1}{L})^x + p_0 \left[ \frac{1-L}{x} - (1 - \frac{1}{L})^x \left( \frac{1-L}{x} + 1 \right) \right] + p_x \left[ 1 - \frac{1-L}{x} + (1 - \frac{1}{L})^x \frac{1-L}{x} \right].$$

But since in practice both  $L$  and  $x$  are large integers we may write

$$(5) \quad \left(1 - \frac{1}{L}\right)^x = e^{-\lambda}, \quad \text{where} \quad \lambda = \frac{x}{L},$$

and (4) then becomes

$$(6) \quad A_x = \alpha A_0 + \beta p_0 + \gamma p_x$$

where

$$(7) \quad \alpha = e^{-\lambda}, \quad \beta = \frac{1 - e^{-\lambda}}{\lambda} - e^{-\lambda}, \quad \gamma = 1 - \frac{1 - e^{-\lambda}}{\lambda}.$$

Tables of  $\alpha$ ,  $\beta$  and  $\gamma$  have been computed for the interval — rate-of-turnover,  $\lambda$ . With the aid of these, tables 1 and 2 are readily extended. A slight difficulty is encountered in determining the acquisition average at an initial point. At the outset it is necessary to assume two initial acquisition averages, one equal to the “high” at some point in the past, the other equal to the corresponding “low.” The true acquisition average certainly lies between these two limits. It is necessary to start computations sufficiently far before the date of the first desired acquisition average so that the two series derived respectively from the highs and lows will converge to a single average. The length of the past experience period required will depend upon the rate of turnover of the stock. The activity in grains is frequently so great that the two series will converge over a period of two weeks.

I wish to point out emphatically that this acquisition average is an average and nothing more. Like any other average its value depends largely upon the ability of the individual using it. Although the use of this average might prove of value to an investor, it can not rightly be said that this is a forecasting formula. I doubt the existence of any valid method of forecasting—mathematical or otherwise. The acquisition average merely measures *secondary* phenomena, and provides a tool for recogniz-

ing an unfavorable condition that might very easily be changed into a favorable situation by any one of numerous causes. Thus, if the market quotation is greater than the acquisition average, it follows that the average owner of the stock in question has a "paper profit." Moreover, since a sale is made when the owner of the stock and the prospective purchaser can agree on a price, and because of the peculiar psychology usually affecting one possessing a paper profit, the excess of the market price over the acquisition average tends through bidding to increase both prices and acquisition averages. This vicious circle carries prices too far in either direction until some "impressed force" changes the trend abruptly.

Since the price of a security at a given time depends upon the status of the entire market as well as the intrinsic value of that security, it follows that a general average for the acquisition figures for a number of the "market leaders" would probably be of value to certain investors. In fact, any of the popular market averages can be accompanied by corresponding acquisition averages.

Since in many cases fifty percent of the stock is kept to protect control, it is evident that one might be justified in using one-half the share units listed for the value of  $L$  in formula (5) for  $\lambda$ . Again, if one desires to investigate the status of the group operating on margins, the amount of the "floating stock" and the brokers' loans must be taken into consideration in determining  $\lambda$ .

In conclusion let me point out that under the most favorable conditions our method of determining the acquisition average can do no more than a 100% successful questionnaire inquiring of stockholders the price at which each share was purchased. Of course all stockholders would not give such information if they could, and couldn't if they would.

# ON A NEW METHOD OF COMPUTING NON- LINEAR REGRESSION CURVES\*

By

WALTER ANDERSSON,  
*fil. dr, Stockholm.*

In a memoir published in this journal in February 1930<sup>1</sup> Professor S. D. Wicksell pointed out that the well-known Pearson method<sup>2</sup> of computing skew regression curves by adopting the principle of least squares can be simplified, and in some direction generalized, by inserting some assumption concerning the distribution function of the population studied. After some remarks on the subject as advanced in the said memoir the problem was presented to me by Professor Wicksell. The results obtained by me as regards this problem were published as a part of my doctor thesis.<sup>3</sup> In the course of the official ventilation of my thesis Professor Wicksell made some interesting remarks concerning the relations between my solution and the general Pearson solution. His suggestion has led me to take up this special problem, which will be considered in the following lines.

I. We consider a bi-variate distribution and denote the variables  $x$  and  $y$ . The distribution function—for the sake of sim-

---

\* From the Statistical Institution of the University of Lund, Sweden.

<sup>1</sup> S. D. Wicksell, Remarks on Regression.

<sup>2</sup> Karl Pearson, On the General Theory of Skew Correlation and Non-Linear Regression; Mathematical Contributions to the Theory of Evolution / *Drap. Comp. Res. Mem., Biom. Ser. II*, 1905.

<sup>3</sup> Walter Andersson, Researches into the Theory of Regression, chapters IV-VI, / *Kungl. Fysiografiska Sällskapets Handlingar*, N. F. Bd. 43, Nr. I; also as *Meddelande från Lunds Observatorium*, Ser. II, Nr. 64 /

plicity being supposed discontinuous—may be

$$(1) \quad z = F(x, y),$$

so that

$$(2) \quad \sum_x \sum_y F(x, y) = 1.$$

Let  $g(x)$  be the regression function of  $y$  on  $x$ . Thus

$$(3) \quad \bar{y}_x = g(x),$$

where  $\bar{y}_x$  denotes the mean value of the dependent variate  $y$  for a fixed value of the independent variate  $x$ . Consequently we have

$$(4) \quad g(x) = \frac{\sum_y y \cdot F(x, y)}{\sum_y F(x, y)}.$$

We further observe that the marginal distribution of  $x$  is

$$(5) \quad f(x) = \sum_y F(x, y).$$

Expanding the regression function in the series of Tchebycheff we put

$$(6) \quad g(x) = \alpha_0 \cdot \psi_0(x) + \alpha_1 \cdot \psi_1(x) + \alpha_2 \cdot \psi_2(x) + \dots,$$

where  $\psi_i(x)$  are polynomials of the  $i^{\text{th}}$  orders, fulfilling the following condition of orthogonality

$$(7) \quad \sum_x f(x) \cdot \psi_i(x) \cdot \psi_j(x) = 0, \quad \text{for } i \neq j,$$

and

$$(8) \quad \sum_x f(x) \cdot [g(x) - \alpha_0 \cdot \psi_0(x) - \alpha_1 \cdot \psi_1(x) - \dots - \alpha_h \cdot \psi_h(x)]^2 = \text{Min.}$$

From (7) and (8) it may be shown that the expansion by Tchebycheff carried to some order gives the same approximate expression for the regression as obtained by fitting a parabola of the same order to the mean values of  $y$  for every value of  $x$ ,

\* Tchebycheff, Collected Works, Vol. I, pp. 203-230.

each observation being allotted a weight proportional to the number of individuals possessing the value of  $x$  in question. Thus, by using the series of Tchebycheff in treating the regression problem we have as a matter of fact applied the same method of describing the regression as applied by Yule<sup>5</sup> and Pearson.<sup>6</sup>

We observe that using the series of Tchebycheff we gain the advantage of being able to perform the graduation successively for the higher orders. With respect to this circumstance I have used the notation *successive regression coefficients* for the coefficients  $\alpha_i$  of (6).

Working out the solution for these coefficients we obtain from (7) and (8),

$$(9) \quad \alpha_i = \frac{\sum_x f(x) \cdot \psi_i(x) \cdot g(x)}{\sum_x f(x) \cdot [\psi_i(x)]^2},$$

the polynomials  $\psi_i(x)$  being determined from (7).

The successive regression coefficients—except  $\alpha_0$ —have been shown / see W. Andersson, Op. cit., pp. 14-15 / to be independent of the zero-values of the variables, and in some cases they are found to stand in simple relations to the well-known semi-invariants of Thiele.<sup>7</sup> Especially when the distribution is assumed to be generated according to the *hypothesis of elementary errors* the semi-invariants of Thiele and the successive regression coefficients are closely related. In this respect the denomination *semi-invariant regression coefficients* may be suggested for the coefficients  $\alpha_i$ . The values of these coefficients ought to be derived in all more exhaustive studies of curved regression lines.

<sup>5</sup> G. U. Yule, On the Significance of Bravais' Formulae for Regression, &c., in case of Skew Correlation / Proc. Roy. Soc., Vol. 60, pp. 477-489, 1897 /.

<sup>6</sup> Pearson, Op. cit.

<sup>7</sup> T. N. Thiele, Theory of Observations, London 1903, p. 24. / See also Annals of Mathematical Statistics, Vol. II, pp. 165-307, where this work of Thiele is reprinted /.

We introduce the *moments*,  $\nu'_{ij}$ , of the distribution. Taking these about any point we have

$$(10) \quad \nu'_{ij} = \sum_x \sum_y x^i \cdot y^j \cdot F(x, y).$$

If we observe that

$$(11) \quad \sum_x f(x) \cdot x^h \cdot g(x) = \sum_x \sum_y x^h \cdot y \cdot F(x, y),$$

it is immediately seen from (9) that the coefficients  $\alpha_i$  can be expressed as linear functions of the "mixed" moments  $\nu'_{01}$ ,  $\nu'_{11}$ ,  $\nu'_{21}$  up to  $\nu'_{h1}$ , all other quantities being dependent on the marginal moments of  $x$  alone.

This solution may shortly be summed up. For a fuller discussion I refer to the cited memoir by the writer.

We write

$$(12) \quad \psi_i(x) = x^i + e_{i, i-1} x^{i-1} + e_{i, i-2} x^{i-2} + \dots + e_{i, 1} x + e_{i, 0}.$$

Let  $\Delta^{(i)}$  be the following determinant of the marginal moments of  $x$ ,

$$(13) \quad \Delta^{(i)} = \begin{vmatrix} 1 & \nu'_{10} & \nu'_{20} & \nu'_{h0} \\ \nu'_{10} & \nu'_{20} & \nu'_{30} & \nu'_{h+1,0} \\ \nu'_{h0} & \nu'_{h+1,0} & \nu'_{h+2,0} & \nu'_{2h,0} \end{vmatrix}$$

and  $\Delta_{hL}$  be its sub-determinant obtained by cutting out the  $(h+1)$ th row and the  $(L+1)$ th column and multiplying by  $(-1)^{h+1}$ . Then we have

$$(14) \quad e_{ij} = \frac{\Delta_{ij}^{(i)}}{\Delta_{ii}^{(i)}},$$

and

$$(15) \quad \alpha_i = \frac{\Delta^{(i-1)}}{\Delta^{(i)}} \left[ \nu'_{i1} + e_{i, i-1} \nu'_{i, i-1} + \dots + e_{i, 1} \nu'_{i1} + e_{i, 0} \nu'_{01} \right],$$

or, using the "standardized" variables

$$(16) \quad \xi = \frac{x - m_1}{\sigma_1}, \quad \eta = \frac{y - m_2}{\sigma_2},$$

( $m = \text{mean}$ ,  $\sigma = \text{dispersion}$ )

and introducing the coefficients

$$(17) \quad g_{i,0} = \varepsilon_{i,1} - r \varepsilon_{i+1,0}$$

where  $\varepsilon_{i,j}$  stands for the "standardized" moments and  $r$  is the usual Galton coefficient of correlation, we have / W. Andersson, Op. cit., p. 16 /

$$(18) \quad \alpha_i = \frac{\Delta^{(i-1)}}{\Delta^{(i)}} \left[ g_{i,0} + e_{i,i-1} g_{i-1,0} + \dots + e_{i,2} g_{3,0} \right].$$

The relations between the successive or the semi-invariant regression coefficients  $\alpha_i$  and the coefficients of the graduation parabolas as written in their usual forms are easily obtained. Taking the parabola of the  $p^{\text{th}}$  order

$$(19) \quad \bar{y}_x = a_0^{(p)} + a_1^{(p)} x + a_2^{(p)} x^2 + \dots + a_p^{(p)} x^p,$$

we have, indeed, / Op. cit., p. 17 /

$$(20) \quad \begin{aligned} a_0^{(p)} &= \alpha_0 + e_{1,0} \alpha_1 + e_{2,0} \alpha_2 + \dots + e_{p,0} \alpha_p \\ a_1^{(p)} &= \alpha_1 + e_{2,1} \alpha_2 + \dots + e_{p,1} \alpha_p \\ a_2^{(p)} &= \alpha_2 + \dots + e_{p,2} \alpha_p \\ &\vdots \\ a_p^{(p)} &= \alpha_p \end{aligned}$$

The coefficients  $e_{i,j}$  are the same as those defined by (14).

2. Starting with the general solution just indicated we may



proceed further into the matter. It will be seen that some new problems are met with in applying the general method to actual statistics.

Taking account of the fact that the solution only involves the moments of the distribution, we can free ourselves from any assumptions as regards the distribution function itself. The required moment values may then be directly computed from the observed frequencies. This way of solving the problem leads to the method advanced by Pearson in his treatises on this subject. The solution evidently gives a least squares graduation to the observed array means when the weights of each mean value are proportional to the observed frequencies in the corresponding arrays.

This method may be the most straight-forward one, but it is, however, by no means the simplest, nor the most efficient one. Considering the fact that the term of the  $i^{th}$  order of the parabola contains moments up to the  $2i^{th}$  order, we immediately conclude that the arithmetical work would rise to a considerable amount, and, with growing moment order be more and more in vain, as a consequence of the rapidly increasing sampling errors of the computed moment values. Some other ways to treat the problem must be sought for in order to eliminate these difficulties.

A first outline of a new method was suggested by S. D. Wicksell in the year 1930 / Wicksell, Op. cit. /. Starting with the general solution Wicksell pointed out that some well-known rules of Thiele as regards the determination of moments of high orders were directly applicable in the computation of high order regression parabolas. The rules of Thiele referred to may be formulated in the following way / Thiele, Op. cit., p. 24 /:

*To obtain the first semi-invariants, or moments, rely entirely on computations. To obtain the intermediate semi-invariants rely partly on computations, partly on theoretical considerations. But to obtain the higher semi-invariants rely entirely on theoretical considerations.*

Professor Wicksell's suggestion was that instead of the higher marginal moments, involved in the least squares expressions for the regression coefficients, should be inserted the moments of a suitably chosen frequency function (with a limited number of parameters), fitted to the marginal distribution of the independent variate.

The method indicated was then more thoroughly studied by the writer of these lines / *Op. cit.* /. The solution obtained along these lines was in detail worked out, and it was also tested as regards its practical usefulness in dealing with actual statistics. Especially by use of the Pearson types of frequency functions very simple expressions, successfully applicable within a large domain of actual statistics, were deduced.

An important advantage of this method / as well as of the Pearson method / of computing high order regression parabolas may be noticed. As the regression coefficients have been expressed as functions of the moments only—in the method elaborated by the author only of those of low orders—the influence of “grouping” may be accounted for by correcting the computed moment values in this respect. For this purpose suitable correction formulas are available, as for instance the well-known ones given by Sheppard. Experience has convinced me that at the ends of the regression curves, at least, the effect of grouping can displace the computed curve in a considerable manner, so that in many cases some attention must be paid to these circumstances.

It is, however, to be remembered that the solution obtained by applying Wicksell's proposition does not give a strict least squares graduation to the observed array means, as a consequence of the fact that the theoretical values of the high order moments always in some degree differ from the directly computed ones. From this it is evident that some care must be taken in choosing the hypothesis as regards the marginal distribution of the independent variate. It may be remarked, however, that these circum-

stances cause very little practical difficulty on account of the much refined theory of uni-variate distributions.

The discrepancy between a least squares solution and the solution as obtained by applying the method as advanced by the author may, as pointed out to me by Professor Wicksell in the course of the official ventilation of my thesis, be removed by an adjustment by which the latter solution is turned into a strict least squares solution. This problem will be considered in the following paragraph, and at the same time we shall get an opportunity to study the hypothetical assumptions applied before from a somewhat different point of view.

3. We consider the expression (8). Before we have from this condition worked out the general least squares solution in assuming  $f(x)$  to be the true marginal distribution (5) and  $g(x)$  to be the true regression function and then 1 / the directly computed moment values were inserted in the general solution / Pearson's method /, or 2 / the moment values required were determined in accordance with the rules indicated in § 2 / method elaborated by the writer /. Now we shall directly imply in (8) our working hypothesis concerning the marginal distribution of  $x$ . Let the hypothetical  $x$ -marginal distribution function be  $\omega(x)$ . The solution is then to be deduced from the following condition

$$(21) \quad \sum_x \omega(x) \cdot [g(x) - \alpha_0 \psi_0(x) - \alpha_1 \psi_1(x) - \dots - \alpha_n \psi_n(x)]^2 = Min.$$

It is immediately clear that, in this way, we always get a strict least squares solution with respect to the distribution function  $\omega(x)$  whatever the form of  $g(x)$  may be.

In fact, the functions  $f(x)$  and  $g(x)$  are totally independent of one another, and, as is seen from (8), the distribution function  $f(x)$  enters into the expansion of Tchebycheff for the regression function  $g(x)$  only as a weight function which determines the weights to be allotted to the regression means in grad-

uating the values of these by means of this series carried to a certain order, or, what is the same, by means of a parabola of the same order, the coefficients of which are determined according to the principle of least squares. Then it is clear that for practical purposes it is not necessary to derive the exact form of  $f(x)$  in performing the expansion. (6). The hypothetical distribution function  $\omega(x)$  would be expected to give a satisfying result as soon as  $\omega(x)$  in its main characteristics corresponds with the true distribution function  $f(x)$ .

I am going to work out the detailed solution for the following two usual forms of  $\omega(x)$ :

A/ Normal Error Function,

$$\omega(\xi) = \frac{1}{\sqrt{2\pi}} e^{-\xi^2/2}$$

(22)

$$\xi = \frac{x - m_1}{\sigma_1},$$

B/ Pearson Type-III Function,

$$\omega(\xi) = C (\xi + \alpha)^{\beta-1} \cdot e^{-\alpha \xi}$$

(23)

$$\xi = \frac{x - m_1}{\sigma_1}; \quad \alpha = -\frac{1}{S}; \quad \beta = \alpha^2.$$

$S$  is the *skewness*, or

$$S = -\frac{\xi_{30}}{Z}.$$

(24)

In both cases the expressions for the terms of the series of Tchebycheff will be found to be very simple.

At first considering the polynomials  $\psi_i(x)$  we are to have in accordance with (7), the distribution function being continuous,

$$(25) \quad \int_{-\infty}^{\infty} dx \cdot \omega(x) \cdot \psi_i(x) \cdot \psi_j(x) = 0 \quad (i \neq j).$$

From this expression it may be concluded that the polynomials  $\psi_i(x)$  are in case A/ the polynomials of Hermite, and in case B/ those of Laguerre. Both these kinds of polynomials are of well-known forms, and consequently the values of the  $e$ -coefficients as defined by (12) may easily be derived from propositions about these polynomials.

For the successive coefficients we have according to (9) the following expression

$$(26) \quad \overline{\alpha}_i = \frac{\int_{-\infty}^{\infty} dx \cdot \omega(x) \cdot \psi_i(x) \cdot g(x)}{\int_{-\infty}^{\infty} dx \cdot \omega(x) \cdot [\psi_i(x)]^2}.$$

Taking account of (13), (14), and (15) and introducing the notation

$$(27) \quad \overline{\nu}_{h1} = \int_{-\infty}^{\infty} dx \cdot \omega(x) \cdot x^h \cdot g(x),$$

we obtain

$$(28) \quad \overline{\alpha}_i = \frac{\overline{\Delta}^{(i-1)}}{\overline{\Delta}^{(i)}} \left[ \overline{\nu}_{i1} + \overline{e}_{i,i-1} \overline{\nu}_{i-1,1} + \dots + \overline{e}_{i1} \overline{\nu}_{11} + \overline{e}_{i0} \overline{\nu}_{01} \right],$$

or, introducing the corresponding "standardized" moments  $\overline{\xi}_{ij}$  and putting

$$(29) \quad \overline{g}_{i0} = \overline{\xi}_{i1} - \overline{\xi}_{11} \overline{\xi}_{i+1,0}$$

we get

$$(30) \quad \overline{\alpha}_i = \frac{\overline{\Delta}^{(i-1)}}{\overline{\Delta}^{(i)}} \left[ \overline{g}_i + \overline{e}_{i,i-1} \overline{g}_{i-1,0} + \dots + \overline{e}_{i2} \overline{g}_{30} \right].$$

The coefficients  $\frac{\overline{\Delta}^{(i-1)}}{\overline{\Delta}^{(i)}}$  and  $\overline{e}_{ij}$  are determined by (14) and (15) when the moment values

$$(31) \quad \bar{\nu}_{h0} = \int_{-\infty}^{\infty} dx \cdot \omega(x) \cdot x^h,$$

respectively

$$(32) \quad \bar{\epsilon}_{h0} = \int_{-\infty}^{\infty} d\xi \cdot \omega(\xi) \cdot \xi^h,$$

are inserted in the determinants.

In the cases here considered we get very simple expressions for these determinants. We have, indeed, / W. Andersson, Op. cit., pp. 88 and 123 / in case of normal distribution

$$(33) \quad \frac{\bar{\Delta}^{(l-1)}}{\bar{\Delta}^{(l)}} = \frac{1}{l!},$$

and in case of Pearson type III distribution

$$(34) \quad \frac{\bar{\Delta}^{(l-1)}}{\bar{\Delta}^{(l)}} = \frac{1}{l! \prod_{h=1}^{l-1} (1+hS^2)}.$$

The values of the  $e$ -coefficients necessary for the computation of the terms of the series of Tchebycheff up to the fifth order are given in the following exposition.

NOR.	$e_{i,j}$	Type III	NOR.	$e_{i,j}$	Type III.
0	$e_{10}$	0	0	$e_{43}$	12S
0	$e_{21}$	2S	-6	$e_{42}$	6(6S <sup>2</sup> -1)
-1	$e_{20}$	-1	0	$e_{41}$	4S(6S <sup>2</sup> -7)
			3	$e_{40}$	-3(6S <sup>2</sup> -1)
0	$e_{32}$	6S	0	$e_{54}$	20S
-3	$e_{31}$	3(2S <sup>2</sup> -1)	-10	$e_{53}$	10(12S <sup>2</sup> -1)
0	$e_{30}$	-4S	0	$e_{52}$	20S(12S <sup>2</sup> -5)
			15	$e_{51}$	5(24S <sup>4</sup> -46S <sup>2</sup> +3)
			0	$e_{50}$	-8S(12S <sup>2</sup> -5)

In order to derive the expressions for the computation of the moment quantities  $\bar{\nu}_h$ , or  $\bar{\epsilon}_h$ , we denote the class-breadth by  $\bar{\omega}$  and the observed mean value of  $y$  in the  $p^{\text{th}}$  array of  $x$  by  $\bar{y}_{x_p}$ . The values of  $\bar{\nu}_h$  are then given by the following formula

$$(35) \quad \bar{\nu}_h = \sum_p I_{x_p} \cdot x_p^h \cdot \bar{y}_{x_p},$$

where

$$(36) \quad I_{x_p} = \int_{x_p - \frac{\bar{\omega}}{2}}^{x_p + \frac{\bar{\omega}}{2}} dx \cdot \omega(x).$$

The computation is easily performed as soon as the function

$$(37) \quad Q(x) = \int_{-\infty}^x dx \cdot \omega(x)$$

is known. In either case we have access to suitable tables of this function. For the Pearson type III function the "Tables for the incomplete  $\Gamma$ -function, edited by Karl Pearson" are to be used.

4. We will now make some general remarks concerning the relations between the different methods of computing regression parabolas touched upon in the preceding lines. We start with the general condition (8) for the determination of the coefficients:

$$\sum_x f(x) \cdot [g(x) - \alpha_0 \psi_0(x) - \alpha_1 \psi_1(x) - \dots - \alpha_h \psi_h(x)]^2 = \text{Min.}$$

It is seen that the expansion is determined by the marginal distribution function  $f(x)$  and the regression function  $g(x)$ . If  $f(x)$  and  $g(x)$  are not the true functions of the population but the functions corresponding to the actual sample, the solution will give the sampling values of the coefficients. This is the solution advanced by Pearson, and consequently in his method no graduation of the data is performed in order to smooth out the

influence of sampling irregularities on the values of the coefficients. Without any further considerations it is clear that methods which include an adjustment of the data in this respect are desirable. The problem is analogous with that occurring in the general theory of distributions. Among other facts of great importance that speak in favour of using mathematical functions for the description of distributions one is that we in this way are able to eliminate in some degree the accidental irregularities. When the regression is described by the series of Tchebycheff the smoothing process is evidently performed firstly by graduating the regression means by a parabola, and secondly by adjusting the parabola coefficients for the accidental irregularities. This latter adjustment has been accounted for by the two methods treated by the author. When using the rules of § 2 as principle for this adjustment the smoothing process is applied to the moment values involved in the general solution for the coefficients, and in the methods indicated in the preceding paragraphs we have used a weight function which is to be considered as a graduation of the observed marginal distribution of the independent variate.

As mentioned before we do not get a strict least squares solution when applying the rules of § 2. This is, however, of little practical importance, but it remains to see in what manner this solution is to be modified in order to become a least squares graduation of the observed array means.

When applying the rules of § 2 the product moments are computed from the following expression

$$(38) \quad \nu'_{h1} = \frac{1}{N} \sum_p n_{x_p} \cdot x_p^h \cdot \bar{y}_{x_p},$$

where  $N$  is the total number of observations and  $n_{x_p}$  the number of observations in the  $p^{th}$  array of  $x$ . We suppose that the graduation is to be based on directly computed moment values



up to the  $h^{th}$  order,  $h$  usually not being greater than six, in accordance with the rules of Thiele. The values of the marginal moments of the independent variate up to the  $h^{th}$  order indicate the distribution function  $f(x, \nu'_{10}, \dots, \nu'_{h0})$ , which function is chosen as the theoretical distribution function determining the values of the marginal moments of orders above the  $h^{th}$ . A strict least squares solution with respect to the distribution function  $f(x, \nu'_{h0})$  may be worked out according to the formulas given in this memoir by taking  $\omega_x = f(x, \nu'_{10}, \dots, \nu'_{h0})$ . In this case we have for the product moments the following values

$$(39) \quad \bar{\nu}'_{h1} = \sum_p I_{x_p} \cdot x_p^h \cdot \bar{y}_{x_p},$$

where

$$(40) \quad I_{x_p} = \int_{x_p - \frac{\omega}{2}}^{x_p + \frac{\omega}{2}} dx \cdot f(x, \nu'_{10}, \dots, \nu'_{h0}).$$

Subtracting (39) from (40) we get

$$(41) \quad \Delta \nu'_{h1} = \sum_p \left( I_{x_p} - \frac{n_{x_p}}{N} \right) \cdot x_p^h \cdot \bar{y}_{x_p},$$

which consequently are the corrections to be added to the directly computed values of the product moments of the solution worked out in accordance with the rules of § 2, in order to obtain a strict least squares solution.

These corrections are easily computed as soon as the integrals  $I_{x_p}$  are determined. This task, however, would in some cases be somewhat arduous. If the general Pearson theory of frequency is applied we must sometimes resort to mechanical quadrature formulas.

a remark concerning the correction of the grouping of the moments  $\bar{U}_{hi}$ . According to the method of computing these characteristics we may regard them as mixed moments of a distribution having as its  $x$ -marginal distribution the function  $\omega(x)$ , the regression means being the observed ones. Thus we evidently can apply the usual methods of correcting computed moment values for the effect of grouping. By using the formulas of Shepard we have to observe, however, that the moments involved in these formulas must be referred to the supposed semi-theoretical distribution.

5. *Numerical Illustrations.* In order to illustrate the application to observed data of the consideration above I have numerically treated a few populations—representative ones in that they are examples of correlation distributions of different degrees of skewness.

*Example I.* Case of slightly skew correlation. Pearson's example B. Example I:2 and II:2 of the cited memoir of the author. Population: Correlation between age and height of head in 2272 girls.

$/x = \text{age}; y = \text{height of head} /$

$$\begin{aligned} x_0 &= 12.5 \text{ yrs.} & \bar{x} - x_0 &= +.2007 & \bar{\omega}_x &= 1 & \xi &= .3263 x' - .0655 \\ y_0 &= 125.25 \text{ mm.} & \bar{y} - y_0 &= -.6017 & \bar{\omega}_y &= 2 & \eta &= .2895 y' + .1742. \end{aligned}$$

As regards the moment values I refer to the memoir of Pearson. These indicate that the marginal distribution of  $x$  may approximately be represented by the normal curve. Thus I take for  $\omega_x$  the normal function.

We have to calculate the product moments  $\bar{U}_{11}$ ,  $\bar{U}_{21}$ , and  $\bar{U}_{31}$ .

These computations may be performed by using the following scheme. The different values are derived from the correlation table given in Pearson's memoir.

The values of  $\xi$  correspond to the class ranges.

$x'$	$\bar{y}'$	$\frac{n_x}{N}$	$\xi$	$I_x$	$I_x - \frac{n_x}{N}$	$x' \cdot \bar{y}'_x$	$x'^2 \cdot \bar{y}'$	$x'^3 \cdot \bar{y}'$
-9	-5.000	.0004	-3.165	.0015	.0011	45.000	-405.000	3645.000
-8	-4.143	.0031	-2.831	.0037	.0006	33.144	-265.152	2121.216
-7	-3.889	.0079	-2.513	.0084	.0005	27.223	-190.561	1333.927
-6	-3.075	.0176	-2.186	.0170	-.0006	18.450	-110.700	664.200
-5	-2.474	.0335	-1.860	.0311	-.0024	12.370	-61.850	309.250
-4	-1.808	.0550	-1.534	.0511	-.0039	7.232	-28.928	115.712
-3	-1.763	.0779	-1.208	.0756	-.0023	5.289	-15.867	47.601
-2	-1.217	.1034	-0.881	.1003	-.0031	2.434	-4.868	9.736
-1	-1.054	.1149	-0.555	.1199	.0050	1.054	-1.054	1.054
0	-0.680	.1360	-0.229	.1296	-.0064	0.000	-0.000	0.000
1	-0.194	.1158	0.098	.1238	.0080	-0.194	-0.194	-0.194
2	0.232	.0871	0.424	.1106	.0235	0.464	0.928	1.856
3	0.453	.0942	0.750	.0858	-.0084	1.359	4.077	12.231
4	0.642	.0713	1.077	.0605	-.0108	2.568	10.272	41.088
5	0.832	.0418	1.403	.0384	-.0034	4.160	20.800	104.000
6	0.885	.0268	1.729	.0218	-.0050	5.310	31.860	191.160
7	2.154	.0057	2.055	.0115	.0058	15.078	105.546	738.822
8	-0.714	.0031	2.382	.0052	.0021	-5.712	-45.696	-365.568
9	0.625	.0035	2.708	.0022	-.0013	5.625	50.625	455.625
10	0.000	.0009	3.034	.0008	-.0001	0.000	0.000	0.000

We get

$$\frac{\sum I_x x' y_x'}{\sum I_x} = 2.9879, \quad \frac{\sum I_x x'^2 y_x'}{\sum I_x} = -6.6564, \quad \frac{\sum I_x x'^3 y_x'}{\sum I_x} = 75.5197$$

and from these values

$$\bar{U}_{11} = 3.1123, \quad \bar{U}_{21} = -2.2376, \quad \bar{U}_{31} = 79.2110.$$

Sheppard's corrections for grouping have been applied.

For the corresponding standardized moments we obtain the following values:

$$\bar{\varepsilon}_{11} = 0.2941, \quad \bar{\varepsilon}_{21} = -0.0689, \quad \bar{\varepsilon}_{31} = 0.8000.$$

This leads to the following values of the  $\bar{g}$ -coefficients defined by (29):

$$\bar{g}_{30} = -0.0689, \quad \bar{g}_{40} = -0.0823.$$

The values of the successive regression coefficients then become

$$\bar{\gamma}_1 = 0.2941, \quad \bar{\gamma}_{21} = -0.0345, \quad \bar{\gamma}_3 = -0.0127.$$

Comparing these different values with the uncorrected ones we find

$$\begin{array}{ll} \bar{\varepsilon}_{11} - \varepsilon_{11} = 0.0000 & \bar{g}_{30} - g_{30} = +0.0021 \\ \bar{\varepsilon}_{21} - \varepsilon_{21} = -0.0086 & \bar{g}_{40} - g_{40} = -0.0337 \\ \bar{\varepsilon}_{31} - \varepsilon_{31} = +0.0511 & \bar{\gamma}_1 - \gamma_1 = +0.0000 \\ \bar{\varepsilon}_{30} - \varepsilon_{30} = -0.0365 & \bar{\gamma}_2 - \gamma_2 = +0.0010 \\ \bar{\varepsilon}_{40} - \varepsilon_{40} = +0.2894 & \bar{\gamma}_3 - \gamma_3 = -0.0056 \end{array}$$

We especially observe that the adjustments of the  $\bar{g}$ -coefficients are smaller than those of the moments of the same orders.

The adjusted coefficients result in the following regression parabola of the third order:

$$\bar{\eta}_\xi = +0.0345 + 0.3352 \xi - 0.0345 \xi^2 - 0.0137 \xi^3.$$

The curve is drawn on diagram 1. For the sake of comparison the graph of the Pearson curve and that obtained by applying the rules of Thiele, the marginal being the normal curve, are given on the same diagram.

*Example II.* Case of moderately skew correlation. Population: Correlation between weight of newborn boy and weight of placenta; material supplied by the Maternity Hospital of Lund, Sweden. Example 2 in S. D. Wicksell: "Correlation Function of Type A, etc." /Kungl. Svenska Vetenskapsakademiens handlingar, Bd 58, Nr. 3/.  $N = 1223$ .

$x$  = weight of boy;  $y$  = weight of placenta/.

$$\begin{aligned} x_0 &= 3350 \text{ gr.} & \bar{\omega}_x &= 300 & \bar{x} - x_0 &= +.4685 & \xi &= .5940 x' - .2783 \\ y_0 &= 630 \text{ gr.} & \bar{\omega}_y &= 80 & \bar{y} - y_0 &= -.5715 & \eta &= .6954 y' + .3974. \end{aligned}$$

The correlation table and the computed moment values are given in the said memoir of Wicksell. For  $\omega_x$  we take the normal function.

Calculating the moments  $\bar{\nu}_{ii}$ , etc. in the same manner as used in the first example we obtain the values,

$$\bar{\nu}_{11} = 1.5540, \quad \bar{\nu}_{21} = 0.3412, \quad \bar{\nu}_{31} = 11.7653$$

which give the following values of the standardized moments:

$$\bar{\varepsilon}_{11} = 0.6420, \quad \bar{\varepsilon}_{21} = 0.0837, \quad \bar{\varepsilon}_{31} = 1.7153.$$

The values are corrected for grouping.

We further get

$$\bar{\varphi}_{30} = +0.0837, \quad \bar{\varphi}_{40} = -0.2106$$

and

$$\bar{\varphi}_1 = +0.6420, \quad \bar{\varphi}_2 = +0.0419, \quad \bar{\varphi}_3 = -0.0351.$$

The values of the adjustments of the different quantities are given below:

$$\begin{array}{ll} \bar{\varepsilon}_{11} - \varepsilon_{11} = -0.0035 & \bar{\varphi}_{30} - \varphi_{30} = -0.0656 \\ \bar{\varepsilon}_{21} - \varepsilon_{21} = +0.0196 & \bar{\varphi}_{40} - \varphi_{40} = -0.0173 \\ \bar{\varepsilon}_{31} - \varepsilon_{31} = -0.2132 & \bar{\varphi}_1 - \varphi_1 = -0.0035 \\ \bar{\varepsilon}_{30} - \varepsilon_{30} = +0.1320 & \bar{\varphi}_2 - \varphi_2 = -0.0328 \\ \bar{\varepsilon}_{40} - \varepsilon_{40} = -0.2870 & \bar{\varphi}_3 - \varphi_3 = -0.0031. \end{array}$$

The correction of  $\varphi_2$  is rather great, but not greater than was to be expected with consideration to the roughness of the fit of the hypothetical marginal distribution function. It is clear that when applying the solution of my previous paper in this example we

should use a type IV curve for the marginal distribution. The unadjusted values of the parabola coefficients are also in this case easily computed, but the calculation of the adjustments by which the solution is turned into a least squares solution would be very laborious.

In order to illustrate the suitability of the several methods I have drawn the following curves on diagram 2: 1/ unadjusted solution, hypothetical marginal distribution being the normal function; 2/ adjusted solution, hypothetical marginal distribution being the normal function; 3/ unadjusted solution, marginal distribution being Pearson's type IV function.

The equation of the second curve is

$$\bar{\eta}_{\xi} = -0.0419 + 0.7473 \xi + 0.0419 \xi^2 - 0.0351 \xi^3.$$

The third curve is undoubtedly best fitted to the data.

*Example III.* Case of extremely skew correlation. The correlation between the age of bachelor and the age of spinster at marriage, Sweden 1911-1920. Example I:4 and II:7 in the cited memoir of the author.  $N = 321908$ .

$$\begin{aligned} /x &= \text{age of spinster}; y = \text{age of bachelor} / \\ x_0 &= 27.5 \text{ yrs.} \quad \bar{x}_0 = 5 \quad \bar{x} - x_0 = -.3131 \quad \xi = .9515 x' + .2929 \\ y_0 &= 27.5 \text{ yrs.} \quad \bar{y}_0 = 5 \quad \bar{y} - y_0 = +.2824 \quad \eta = .8506 y' - .2424. \end{aligned}$$

The moment values as given in the said memoir indicate that we can use Pearson's type III function as hypothetical distribution function for the  $x$ -marginal distribution. From the moment values as computed in the cited memoir we obtain the following values of the constants of this function:

$$\alpha = 1.4312 \quad \beta = 2.0483.$$

It is to be remarked that for our purposes the computation of the constant  $C$  is not necessary.

For the  $c$ -coefficients we get the following values:

$$\begin{aligned} e_{21} &= -1.3974 & e_{32} &= -4.1922 & e_{43} &= -8.3844 \\ e_{20} &= -1.0000 & e_{31} &= -0.0708 & e_{42} &= +11.5752 \\ & & e_{30} &= +2.7948 & e_{41} &= +11.3771 \\ & & & & e_{40} &= -5.7876. \end{aligned}$$

For the unadjusted values of the  $\mathcal{S}$ -coefficients and the successive regression coefficients we further get

$$\mathcal{S}_{30} = +0.1787 \quad \mathcal{S}_{40} = +0.5255 \quad \mathcal{S}_{50} = +2.8763$$

$$\alpha_1 = +0.5535 \quad \alpha_2 = +0.05192$$

$$\alpha_3 = -0.0122669 \quad \alpha_4 = +0.003097.$$

Computing the corresponding adjusted values by use of "Tables of the incomplete  $\sqrt{\cdot}$ -function" we obtain

$$\bar{\mathcal{S}}_{30} = +0.1723 \quad \bar{\mathcal{S}}_{40} = +0.4638 \quad \bar{\mathcal{S}}_{50} = +2.0851$$

$$\bar{\alpha}_1 = +0.5528 \quad \bar{\alpha}_2 = +0.05789$$

$$\bar{\alpha}_3 = -0.014647 \quad \bar{\alpha}_4 = +0.001097$$

Sheppard's corrections have been applied in both cases. The differences between the adjusted and the unadjusted values are

$\bar{\varepsilon}_{11} - \varepsilon_{11} = -0.0007$	$\bar{\varepsilon}_{30} - \varepsilon_{30} = 0.0000$
$\bar{\varepsilon}_{21} - \varepsilon_{21} = -0.0034$	$\bar{\varepsilon}_{40} - \varepsilon_{40} = -0.4762$
$\bar{\varepsilon}_{31} - \varepsilon_{31} = -0.3237$	$\bar{\varepsilon}_{50} - \varepsilon_{50} = -2.0405$
$\bar{\varepsilon}_{41} - \varepsilon_{41} = -1.4548$	$\bar{\alpha}_1 - \alpha_1 = -0.0007$
$\bar{\mathcal{S}}_{30} - \mathcal{S}_{30} = -0.0064$	$\bar{\mathcal{S}}_{40} - \mathcal{S}_{40} = -0.0617$
$\bar{\mathcal{S}}_{50} - \mathcal{S}_{50} = -0.7912$	$\bar{\alpha}_2 - \alpha_2 = +0.00597$
	$\bar{\alpha}_3 - \alpha_3 = -0.001978$
	$\bar{\alpha}_4 - \alpha_4 = -0.002000.$

The parabolas of the third and the fourth orders are the following ones:

Unadjusted values of the coefficients:

$$\bar{\eta}_{\xi} = -0.0873 + .4848\xi + .1050\xi^2 - .01267\xi^3$$

$$\bar{\eta}_{\xi} = -0.1053 + .5171\xi + .1409\xi^2 - .03864\xi^3 + .003097\xi^4.$$

Adjusted values of the coefficients:

$$\bar{\eta}_{\xi} = -0.0988 + .4729\xi + .1193\xi^2 - .01465\xi^3$$

$$\bar{\eta}_{\xi} = -0.1052 + .4854\xi + .1320\xi^2 - .02384\xi^3 + .001097\xi^4.$$

The graphs are drawn on diagrams 3 and 4.

The results indicated by the few examples treated in this paragraph clearly point out that the Tchebycheff expansion cannot be considered as a least squares graduation of the observed

regression means when the moment values involved in the solution are determined in accordance with the rules laid out in § 2. As regards the practical applicability of such a solution, however, this circumstance is of little importance, because the curve in this case is found to give as good, and sometimes a better representation of the regression than a strict least squares graduation. Further, as the calculation of the moments of the first few orders is often required for other purposes than the determination of the regression curve, the computation of the unadjusted solution in these cases is arithmetically very simple. Not having access to the moment values we may perhaps in some cases consider the direct computation of the adjusted solution as performed in example I to be the simplest method. The adjusting of correctly determined unadjusted solutions would certainly very seldom be of real gain.

Stockholm, September 1933.

S. D. WICKSELL

*Note on Dr. Andersson's Paper.*

In an extensive memoir, *Researches into the theory of Regression*, Dr. W. Andersson has worked out a very simple and widely applicable numerical method of computing curved regressions. The general principle on which this method was founded Dr. Andersson has kindly attributed to me. It was laid out in my paper in the first number of the "Annals" Journal and may be stated as follows: After fitting a suitable univariate frequency function with a limited number of parameters—e.g. the normal curve or one of Pearson's types—to the marginal distribution of the independent variate, the moments of this function—which are all expressible in terms of the parameters—should be used in computing the regression coefficients, instead of the ordinary



values (power means). Of course, when, in fitting the curve, the ordinary moments of lower orders have been used in determining the parameters, this procedure means that the moments of higher orders are theoretically expressed in terms of the moments of lower orders instead of being directly computed.

Applying this device to the ordinary least squares expressions for the regression coefficients, it was clear that a departure from the least square condition took place, but the chances were that this would not harm the result, and the computations would be much simplified. Dr. Andersson's investigation has shown that these expectations were highly justified.

During the official ventilation of the memoir, which was presented as Thesis for the degree of D.Ph., it was agreed that the method ought to be tested by a comparison with a theoretically very similar method in which the least squares condition was retained, although theoretical or semi-empirical weights were introduced instead of the purely empirical weights used in the method of Karl Pearson.

In the present paper Dr. Andersson has taken this question up and he shows that whereas the original (unadjusted) method is numerically simpler in application, it gives practically just as good regression curves as the new, adjusted method. In some cases he even considers the unadjusted solution to be the better one.

By this the incident may seem to be closed. I should, however, like to point out, in a few words, how very straightforward a principle it is, which lies behind this adjusted method.

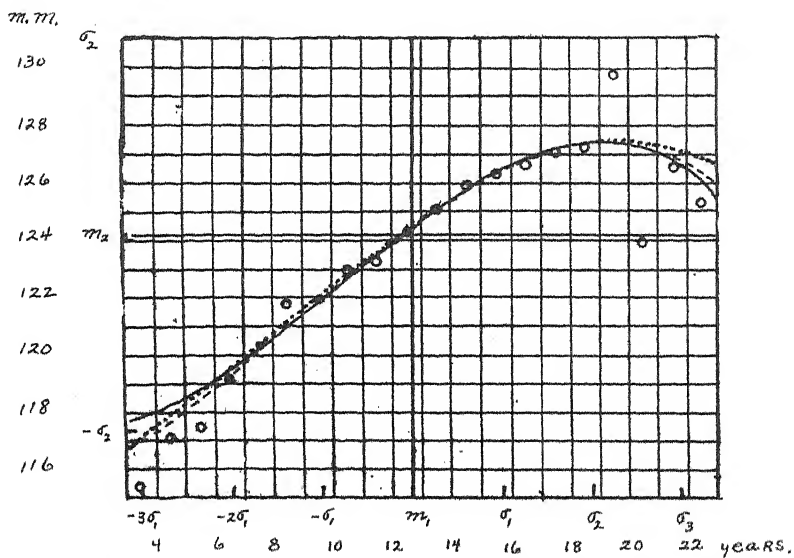
It is simply this: When a correlation table is given, the regression of  $y$  on  $x$  will not be affected by multiplying the frequencies within any  $x$ -array by a constant factor. Hence the following procedure will not affect the regression of  $y$  on  $x$ ; i.e., the process of reducing or adjusting the frequencies in the several  $x$ -arrays so that the marginal sums will be equal to the smoothed frequencies, corresponding to any mathematical curve which has

been fitted to the marginal distribution. Thus, on applying Pearson's ordinary least squares solution to this adjusted table a least squares regression parabola would be obtained in which the marginal moments were those of the smoothed distribution, and also the mixed moments were, although only in a secondary degree, affected by the smoothing of the marginal. It is only in this last respect, i.e. as regards the mixed moments, that this method deviates from the one originally proposed.

In my opinion many curved regressions could be very easily and accurately enough computed by simply smoothing the marginal of the independent variate with a normal curve or, eventually, a Pearson Type III curve, and correspondingly adjusting the array frequencies. This method may work well even if the deviations of the actual distribution from the smoothed distribution are systematical.

Statistical Institute, University of Lund, November 1933.

DIAGRAM 1

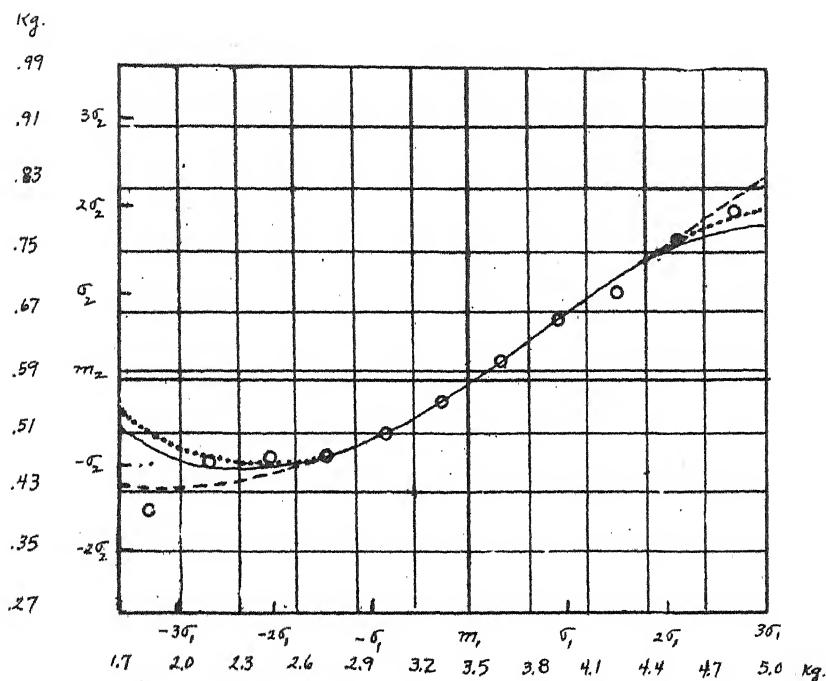
*Cubics*

Unbroken curve: Adjusted solution, the hypothetical marginal distribution being the normal curve.

Dotted curve: Unadjusted solution, the hypothetical marginal distribution being the normal curve.

Dashed curve: Pearson's curve.

DIAGRAM 2

*Cubics*

Unbroken curve: Adjusted solution, the hypothetical marginal distribution being the normal curve.

Dotted curve: Unadjusted solution, the hypothetical marginal distribution being the normal curve.

Dashed curve: Unadjusted solution, the hypothetical marginal distribution being the Pearson Type IV curve.

DIAGRAM 3

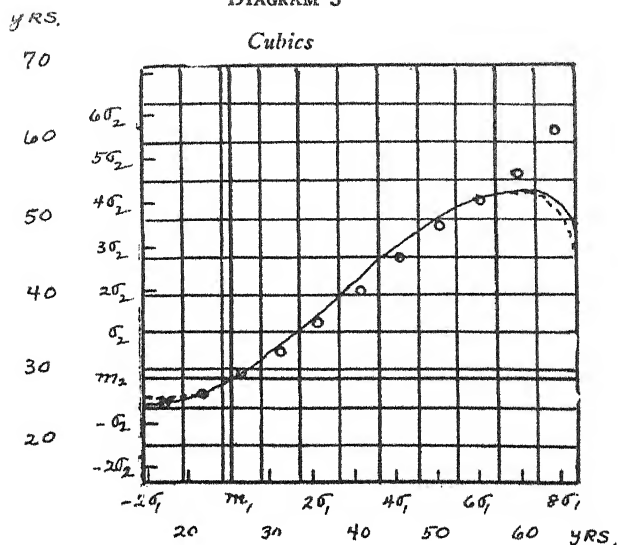
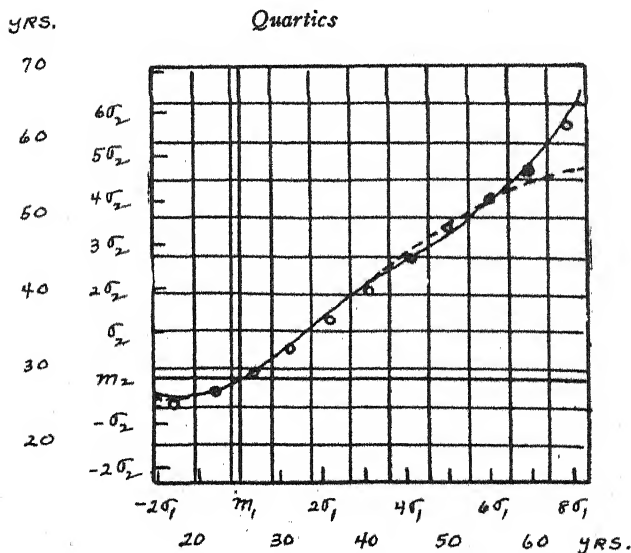


DIAGRAM 4



Unbroken curves: Unadjusted solutions. Dashed curves: Adjusted solutions, the hypothetical marginal distribution being the Pearson Type III curve.

# THE STANDARD ERROR OF ANY ANALYTIC FUNCTION OF A SET OF PARAMETERS EVALUATED BY THE METHOD OF LEAST SQUARES

By

WALTER A. HENDRICKS,  
Bureau of Animal Industry,  
U. S. Department of Agriculture, Washington, D. C.

After fitting a curve to a set of data by the method of least squares, it is occasionally necessary to use the resulting values of some or all of the parameters of the curve in further calculations. Since the estimates of the values of the parameters obtained from a particular set of data are subject to errors of sampling, it follows that the result of any calculation involving those values of the parameters will have a certain standard error. Since the estimated values of the parameters are not independent of each other, the familiar formulas based on the assumption of independence should not be used for the purpose of calculating this standard error from the standard errors of the parameters themselves. The correct approach to the problem involves little more than an application of the methods presented by Schultz (1930) in his excellent paper describing the method of calculating the standard error of a particular function of the parameters, viz., the same function which was used in evaluating the parameters.

Let  $y = \varphi(\lambda_1, \lambda_2, \dots, \lambda_k)$  be an analytic function involving the  $k$  parameters,  $\lambda_i$ . This function may not be linear with respect to the parameters, so that if the parameters are to be evaluated by the method of least squares, we have in the general case a function of the form:

$$(1) \quad y = \varphi(\lambda_1, \lambda_2, \dots, \lambda_k) + \frac{\partial \varphi}{\partial \lambda_1} \Delta \lambda_1 + \dots + \frac{\partial \varphi}{\partial \lambda_k} \Delta \lambda_k + \dots$$

from which the values of the parameters may be obtained by assuming approximate values and calculating the corrections which must be added to obtain the most probable values.

After the values of the parameters have been obtained, let it be required to find the standard error of a new function,  $z = \theta(\lambda_1, \lambda_2, \dots, \lambda_k)$ , involving those values. If  $z$  is an analytic function of the parameters, we have to a close approximation:

$$(2) \quad z = \theta(\lambda_1, \lambda_2, \dots, \lambda_k) + \frac{\partial \theta}{\partial \lambda_1} \Delta \lambda_1 + \frac{\partial \theta}{\partial \lambda_2} \Delta \lambda_2 + \dots + \frac{\partial \theta}{\partial \lambda_k} \Delta \lambda_k.$$

Any error in  $z$ , beyond the insignificant error introduced by the above expansion, will then be due only to errors in  $\Delta \lambda_1, \Delta \lambda_2, \dots, \Delta \lambda_k$ .

Therefore, if

$$(3) \quad f = \frac{\partial \theta}{\partial \lambda_1} \Delta \lambda_1 + \frac{\partial \theta}{\partial \lambda_2} \Delta \lambda_2 + \dots + \frac{\partial \theta}{\partial \lambda_k} \Delta \lambda_k$$

and  $S_z$  and  $S_f$  denote the standard errors of  $z$  and  $f$ , respectively, it is at once apparent that  $S_z = S_f$ .

The values of  $\Delta \lambda_1, \Delta \lambda_2, \dots, \Delta \lambda_k$  may be expressed in terms of the data from which they were evaluated,

$$(4) \quad \begin{cases} \Delta \lambda_1 = \sigma_1 M_1 + \sigma_2 M_2 + \dots + \sigma_n M_n \\ \Delta \lambda_2 = \tau_1 M_1 + \tau_2 M_2 + \dots + \tau_n M_n \\ \Delta \lambda_k = \xi_1 M_1 + \xi_2 M_2 + \dots + \xi_n M_n \end{cases}$$

in which the values,  $M_i$ , represent the  $n$  observed values of the variable,  $y$ ;  $f$  may then be expressed in the form:

$$(5) \quad f = \sum_{i=1}^n \left[ \frac{\partial \theta}{\partial \lambda_1} \sigma_i M_i + \frac{\partial \theta}{\partial \lambda_2} \tau_i M_i + \dots + \frac{\partial \theta}{\partial \lambda_k} \xi_i M_i \right].$$

From the well-known laws of propagation of error and the fact that  $S_z = S_f$ , it follows that

$$(6) \quad S_z^2 = \sum_{i=1}^n \left[ \frac{\partial \theta}{\partial \lambda_1} \sigma_i + \frac{\partial \theta}{\partial \lambda_2} \tau_i + \dots + \frac{\partial \theta}{\partial \lambda_k} \xi_i \right]^2 S_y^2,$$

in which  $S_y$  is the standard error of estimate of  $y$  based on  $n-k$

degrees of freedom. If the right-hand member of equation (6) is expanded, the equation may be written in the form:

$$(7) \quad S_z^2 = \left\{ \left( \frac{\partial \theta}{\partial \lambda_1} \right)^2 [\sigma \sigma] + \left( \frac{\partial \theta}{\partial \lambda_2} \right)^2 [\tau \tau] + \cdots + \left( \frac{\partial \theta}{\partial \lambda_k} \right)^2 [\xi \xi] + \right. \\ \left. 2 \left( \frac{\partial \theta}{\partial \lambda_1} \right) \left( \frac{\partial \theta}{\partial \lambda_2} \right) [\sigma \tau] + \cdots + 2 \left( \frac{\partial \theta}{\partial \lambda_1} \right) \left( \frac{\partial \theta}{\partial \lambda_k} \right) [\sigma \xi] + \right. \\ \left. \cdots + 2 \left( \frac{\partial \theta}{\partial \lambda_2} \right) \left( \frac{\partial \theta}{\partial \lambda_k} \right) [\tau \xi] + \cdots \right\} S_y^2,$$

in which  $[\sigma \sigma] = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2$ , etc.

The values of the sums of the squares and products multiplying the differential coefficients in equation (7) may be obtained from the normal equations formed for the evaluation of  $\Delta \lambda_1, \Delta \lambda_2, \dots, \Delta \lambda_k$ . Let the normal equations be:

$$(8) \quad \begin{aligned} [aa] \Delta \lambda_1 + [ab] \Delta \lambda_2 + \cdots + [al] \Delta \lambda_k &= [aM] \\ [ba] \Delta \lambda_1 + [bb] \Delta \lambda_2 + \cdots + [bl] \Delta \lambda_k &= [bM] \\ \vdots &\vdots \\ [la] \Delta \lambda_1 + [lb] \Delta \lambda_2 + \cdots + [ll] \Delta \lambda_k &= [lM] \end{aligned}$$

If these equations are solved for  $\Delta \lambda_1, \Delta \lambda_2, \dots, \Delta \lambda_k$  by the method of undetermined multipliers, the first is multiplied by an undetermined constant,  $\alpha_1$ , the second by  $\alpha_2$ , etc., and the resulting products are added. The conditions for the solution for  $\Delta \lambda_1$  are:

$$(9) \quad \begin{aligned} [aa] \alpha_1 + [ab] \alpha_2 + \cdots + [al] \alpha_k &= 1 \\ [ba] \alpha_1 + [bb] \alpha_2 + \cdots + [bl] \alpha_k &= 0 \\ [la] \alpha_1 + [lb] \alpha_2 + \cdots + [ll] \alpha_k &= 0. \end{aligned}$$

To solve for  $\Delta \lambda_2$ , equations (8) are multiplied by  $\beta_1, \beta_2, \dots, \beta_k$ , respectively, added, and the following conditions imposed:

$$(10) \quad \begin{aligned} [aa] \beta_1 + [ab] \beta_2 + \cdots + [al] \beta_k &= 0 \\ [ba] \beta_1 + [bb] \beta_2 + \cdots + [bl] \beta_k &= 1 \\ [la] \beta_1 + [lb] \beta_2 + \cdots + [ll] \beta_k &= 0. \end{aligned}$$

To solve for  $\Delta \lambda_k$ , equations (8) are multiplied by  $\omega_1, \omega_2, \dots, \omega_k$ , respectively, added, and the following conditions imposed:



$$\begin{aligned}
 & [aa] \omega_1 + [ab] \omega_2 + \dots + [al] \omega_k = 0 \\
 (11) \quad & [ba] \omega_1 + [bb] \omega_2 + \dots + [bl] \omega_k = 0 \\
 & [la] \omega_1 + [lb] \omega_2 + \dots + [ll] \omega_k = 1
 \end{aligned}$$

It may be proved that:

$$\begin{aligned}
 (12) \quad & \alpha_1 = [\sigma\sigma] \quad \beta_1 = [\tau\sigma] \quad \omega_1 = [\xi\sigma] \\
 & \alpha_2 = [\sigma\tau] \quad \beta_2 = [\tau\tau] \quad \omega_2 = [\xi\tau] \\
 & \alpha_k = [\sigma\xi] \quad \beta_k = [\tau\xi] \quad \omega_k = [\xi\xi].
 \end{aligned}$$

The method of deriving equations (12) is indicated in the well-known text on the method of least squares by Merriman (1907) in which a detailed proof of the fact that  $\beta_2$  is equal to  $[\tau\tau]$  is presented. The other relations may be derived in analogous fashion. It may be observed that  $[\sigma\tau] = [\tau\sigma]$ , etc.

The required quantities to be substituted in equation (7) may, therefore, be calculated by solving the sets of simultaneous equations, (9), (10), and (11).

This completes the solution of the general problem presented in the first part of this paper. Some confusion may arise in regard to the proper application of the methods described above if one or both of the functions,  $y$  and  $z$ , happens to be in a linear form with respect to the parameters. It may be shown that the formulas given will hold in any of these special cases. Although Taylor's theorem may be applied to such functions, such a treatment is superfluous. If either or both of the functions,  $y$  and  $z$ , is linear with respect to the parameters, the expression for  $S_z^2$  is identical with equation (7) even though the linear function, or functions, was not first expanded by Taylor's theorem. Furthermore, if  $y$  is linear with respect to the parameters, the values of the coefficients,  $[\sigma\sigma]$ , etc., in equation (7) will be the same regardless of whether the parameters were evaluated directly or whether  $y$  was first expanded. The latter statement is evident from

an inspection of equations (9), (10), and (11) and a consideration of the law of formation of normal equations.

As an example of the application of the methods presented in this paper to a specific problem, consider a set of data given by Spillman (1933) relating to the yields of potatoes obtained from four plots of ground which had been treated with different amounts of potash.

YIELDS OF POTATOES FROM FOUR PLOTS OF GROUND RECEIVING  
DIFFERENT AMOUNTS OF POTASH

$x$ (Units of $K_2O$ )	$y$ (Bushels of potatoes)
0	91
1	251
2	331
3	381

When a simple exponential equation of the form,  
(13)  $y = A - Be^{-\kappa x}$   
was fitted to this set of data, the most probable values of the parameters,  $A$ ,  $B$ , and  $\kappa$ , were found to be as follows:

$$A = 432.801 \pm 11.637$$

$$B = 341.393 \pm 11.406$$

$$\kappa = 0.6195918 \pm 0.0462871 .$$

The value of the product of the parameters,  $A$  and  $\kappa$ , happens to be of some interest, at least to the author of this paper, since it gives the value of the first derivative of  $y$  with respect to  $x$  at the point where the curve crosses the  $x$ -axis. In the present example it represents the increase in yield, per unit increase in amount of potash applied, which would be possible if certain inhibiting influences, which seem to be proportional to the yield, had no effect. For the particular data under consideration, the value of this product is 268.160.

In order to calculate the standard error of this value, equation (7) was applied as follows:

$$(14) \quad S_z^2 = \left( \kappa^2 [\sigma\sigma] + A^2 [\xi\xi] + 2A\kappa [\sigma\xi] \right) S_y^2,$$

from which the standard error of  $A\kappa$  was found to be equal to  $\pm 13.331$ .

The familiar formula for calculating the standard error of the product of two independent quantities, when employed for the purpose of calculating the standard error of  $A\kappa$  may be written in the form:

$$(15) \quad S_z^2 = \left( \kappa^2 [\sigma\sigma] + A^2 [\xi\xi] \right) S_y^2.$$

Equation (15) gives a value of  $\pm 21.291$  for the standard error of  $A\kappa$ , which deviates considerably from the correct value given by equation (14). The discrepancy is due entirely to the fact that the estimated values of  $A$  and  $\kappa$  are not independent.

#### REFERENCES

- MERRIMAN, MANSFIELD. 1907. *The Method of Least Squares*. John Wiley & Sons, New York.
- SCHULTZ, HENRY. 1930. The standard error of a forecast from a curve. *Jour. Amer. Stat. Assoc.*, 25 (N. S. 17): 139-185.
- SPILLMAN, W. J. 1933. Use of the exponential yield curve in fertilizer experiments. *U. S. D. A. Tech. Bull.* 348.

# TRANSFORMATION OF NON-NORMAL FREQUENCY DISTRIBUTIONS INTO NORMAL DISTRIBUTIONS\*

By

G. A. BAKER

This investigation is undertaken for two reasons: (1) there has been a demand on the part of some statisticians for an analytic method of transforming non-normal distributions into normal distributions; and, (2) a non-normal distribution and the transformation necessary to transform it into a normal distribution serve to specify the distributions in random samples of estimates of the parameters of the original distribution in terms of the distributions of estimates of the parameters of a normal distribution in random samples. In this way valuable approximations to the distributions of the parameters of the original non-normal population may be secured.

## PART I

### TRANSFORMATIONS OF FREQUENCY DISTRIBUTIONS

Consider a non-normal frequency distribution represented by  $f(x) dx$  where the origin is taken at some central point, say the mode, mean, or median, or near one of these points, and the scale is, or approximately is, the standard deviation of the distribution. We seek a function,  $\varphi$ , such that  $f(x) dx$  transformed by the transformation,  $x = \varphi(u)$ , becomes a normal distribution of total area  $\sqrt{2\pi}$ , mean zero and standard deviation unity, i.e.

---

\* Presented at the May, 1932 meeting of the Illinois section of the American Mathematical Association.

$$(1) \quad f[\varphi(u)] \cdot \varphi'(u) du = e^{-\frac{1}{2}u^2} du.$$

In a previous paper<sup>1</sup> expressions similar to (1) were regarded as differential equations which can be solved exactly in certain special cases. In the case of (1) it seems preferable to regard

$$(2) \quad f[\varphi(u)] \varphi'(u) = e^{-\frac{1}{2}u^2}$$

as an identity in  $u$ . If it is assumed that  $f$  and  $\varphi$  are functions that can be represented by Maclaurin's expansions, which is a reasonable assumption regarding  $f$  and  $\varphi$  if  $f$  is near normal, the two members of (2) can be expanded and the coefficients of corresponding powers of  $u$  can be equated thus determining  $\varphi$ .

Suppose that

$$(3) \quad f(x) = \sum_{n=0}^{\infty} A_n x^n$$

$$(4) \quad \varphi(x) = \sum_{n=1}^{\infty} B_n x^n$$

$$(5) \quad \varphi'(x) = \sum_{n=1}^{\infty} n \cdot B_n \cdot x^{n-1}.$$

Then (2) becomes

$$(6) \quad \sum_{m=0}^{\infty} A_m \left[ \sum_{n=1}^{\infty} B_n u^n \right]^m \cdot \sum_{n=1}^{\infty} n \cdot B_n u^{n-1} = 1 - \frac{u^2}{2} + \frac{u^4}{2 \cdot 4} - \frac{u^6}{2 \cdot 4 \cdot 6} + \frac{u^8}{2 \cdot 4 \cdot 6 \cdot 8} - \frac{u^{10}}{2 \cdot 4 \cdot 6 \cdot 8 \cdot 10} + \dots$$

Hence

$$B_1 = \frac{1}{A_0}, \quad B_2 = -\frac{A_1 B_1^3}{2}, \quad B_3 = -\frac{B_1}{6} - A_1 B_1^2 B_2 - \frac{A_2 B_1^4}{3},$$

<sup>1</sup> Transformations of Bimodal Distributions, *Annals of Mathematical Statistics*, Vol. I, No. 4, Nov. 1930.

$$B_4 = -A_1(B_1^2 B_3 + \frac{B_1 B_2^2}{2}) - \frac{A_2 B_1^3 B_2}{2} - \frac{A_3 B_1^5}{4}$$

$$B_5 = \frac{B_1}{40} - A_1(B_1^2 B_4 + B_1 B_2 B_3) - A_2(B_1^3 B_3 + B_1^2 B_2^2) - A_3 B_1^4 B_2 - \frac{A_4 B_1^6}{5}$$

$$\begin{aligned} B_6 = & -A_1(B_1^2 B_5 + B_1 B_2 B_4 + \frac{B_1 B_3^2}{2}) \\ & - A_2(B_1^3 B_4 + \frac{B_1 B_2^3}{3} + 2 B_1^2 B_2 B_3) \\ & - A_3(B_1^4 B_3 + B_1^3 B_2^2) - A_4 B_1^5 B_2 - \frac{A_5 B_1^7}{6} \end{aligned}$$

$$\begin{aligned} B_7 = & -\frac{B_1}{336} - A_1(B_1^2 B_6 + B_1 B_2 B_5 + B_1 B_3 B_4) \\ & - A_2(B_1^3 B_5 + B_1 B_2^2 B_3 + B_1^2 B_3^2 + 2 B_1^2 B_2 B_4) \\ & - A_3(B_1^4 B_4 + \frac{B_1^2 B_2^3}{7} + 3 B_1^3 B_2 B_3) \\ & - A_4(B_1^5 B_3 - \frac{8}{7} B_1^4 B_2^2) - A_5 B_1^6 B_2 - \frac{A_6 B_1^8}{7} \end{aligned}$$

$$\begin{aligned} B_8 = & -A_1(B_1^2 B_7 + B_1 B_2 B_6 + B_1 B_3 B_5 + \frac{1}{2} B_1 B_4^2) \\ & - A_2(B_1^3 B_6 + B_1 B_2^2 B_4 + B_1 B_2 B_3^2 + 2 B_1^2 B_2 B_5 + 2 B_1^2 B_3 B_4) \\ & - A_3(B_1^4 B_5 + \frac{B_1 B_2^4}{4} + 3 B_1^3 B_2 B_4 + \frac{9}{8} B_1^3 B_3^2 + \frac{15}{8} B_1^2 B_2^2 B_3) \\ & - A_4(B_1^5 B_4 + 4 B_1^4 B_2 B_3 + \frac{3}{2} B_1^3 B_2^3) \\ & - A_5(B_1^6 B_3 + \frac{5}{2} B_1^5 B_2^2) \\ & - A_6 B_1^7 B_2 - \frac{A_7 B_1^9}{8} \end{aligned}$$

The corresponding formulas for determining a function to transform a normal distribution of total area  $\sqrt{2\pi}$ , mean at zero and standard deviation of unity, into a given non-normal distribution are as follows. (The  $A_s$  are the coefficients in the expansion of the given non-normal distribution and the  $B_s$  are the coefficients of the transforming function.)

$$B_1 = A_0, \quad B_2 = \frac{A_1}{2}, \quad B_3 = \frac{A_2}{2} + \frac{B_1^3}{6},$$

$$B_4 = \frac{1}{4} A_3 + \frac{1}{4} B_1^2 B_2$$

$$B_5 = \frac{1}{5} A_4 + \frac{1}{2} B_1^2 B_3 + \frac{1}{2} B_1 B_2^2 - \frac{1}{40} B_1^5$$

$$B_6 = \frac{1}{6} A_5 + \frac{1}{2} B_1^2 B_4 + \frac{1}{6} B_2^3 + B_1 B_2 B_3 - \frac{1}{8} B_1^4 B_2$$

$$B_7 = \frac{1}{7} A_6 + \frac{1}{2} B_1^2 B_5 + \frac{1}{2} B_2^2 B_3 + \frac{1}{2} B_1 B_3^2 + B_1 B_2 B_4 \\ - \frac{1}{8} B_1^4 B_3 - \frac{1}{7} B_1^3 B_2^2 + \frac{B_1^7}{336}$$

$$B_8 = \frac{1}{8} A_7 + \frac{1}{2} B_1^2 B_6 + \frac{1}{2} B_2^2 B_4 + \frac{1}{2} B_2 B_3^2$$

$$+ B_1 B_2 B_5 + B_1 B_3 B_4 - \frac{1}{8} B_1^4 B_4$$

$$- \frac{1}{2} B_1^3 B_2 B_3 - \frac{3}{16} B_1^2 B_3^2 + \frac{1}{48} B_1^6 B_2$$

These formulas give very simple results for the expression of the first few terms of the transforming function,  $\varphi$ , in terms of the coefficients of the given function. If the coefficients in the expansion of  $\varphi$  rapidly approach zero so that only a few terms are needed for a good approximation the method outlined should

be effective. Edgeworth<sup>2</sup> has discussed at some length the transformation or "translation" of normal distributions into non-normal distributions and has given several methods of determining the coefficients of the transforming function. The formulas presented here are more simple but their practicability can be demonstrated only by numerical results in special cases. For practical purposes the left-hand member of (6) need represent the right-hand member accurately only in the interval, say  $-2 \leq u \leq 2$ .

## ILLUSTRATION

For example, consider

$$f(x) = .9929 \left(1 + \frac{x}{10}\right)^{.99} e^{-.10x},$$

which is skewed noticeably in the positive direction but which is of a type that approaches a normal distribution as the skewness approaches zero. Then

$$\begin{array}{ll} A_0 = .9929 & B_1 = 1.0072 \\ A_1 = -.1000 & B_2 = .0511 \\ A_2 = -.4887 & B_3 = .0050 \\ A_3 = .0823 & B_4 = -.0080 \\ A_4 = .1142 & B_5 = .0004 \\ A_5 = -.0279 & \end{array}$$

<sup>2</sup> Bowley, A. L.-F. Y. Edgeworth's Contributions to Mathematical Statistics, pp. 65-78.



TABLE I

Comparison of the ordinates of the normal function, function with skewness of .2, and the skewed function transformed by the transformation  $x = 1.0072 u + .0511 u^2 + .0050 u^3 - .0080 u^4$ .

u	Normal curve*	Function with Skew .2†	Transformed skew curve†	Normal minus skew curve	Normal minus transformed skew curve
2.0	.053991	.049243	.0576	.0047	.0036
1.8	.078950	.076810	.0910	.0021	.0120
1.6	.110921	.112956	.1327	.0020	.0118
1.4	.149727	.157043	.1715	.0073	.0218
1.2	.194186	.206951	.2099	.0128	.0157
1.0	.241971	.259120	.2505	.0171	.0085
0.8	.289692	.308958	.2897	.0193	.0058
0.6	.333225	.351538	.3366	.0183	.0033
0.4	.368270	.382453	.3776	.0142	.0093
0.2	.391043	.398583	.3907	.0075	.0004
0.0	.398942	.398859	.3989	.0001	.0000
0.2	.391043	.383157	.3906	.0079	.0005
0.4	.368270	.354545	.3688	.0137	.0005
0.6	.333225	.316273	.3299	.0170	.0033
0.8	.289692	.272360	.2842	.0173	.0055
1.0	.241971	.226714	.2323	.0153	.0097
1.2	.194186	.182641	.1803	.0115	.0139
1.4	.149727	.142563	.1319	.0072	.0178
1.6	.110921	.107939	.0908	.0030	.0202
1.8	.078950	.079354	.0717	.0004	.0073
2.0	.053991	.056702	.0452	.0027	.0088

\* These columns were taken from Luis R. Salvosa's tables, *Annals of Mathematical Statistics*, Vol. I, No. 2, p. 64 et seq.

† These values were calculated by interpolating in the above mentioned tables.

The ordinates of the normal curve,  $f(x)$ , and  $f(x)$  transformed by the transformation determined by the first four  $B$ 's are compared in Table I.

The ordinates of the transformed distribution are much nearer those of the normal curve over an interval that includes seventy-five per cent of the frequency but for the rest of the range considered the agreement is not so good. These facts indi-

cate that more terms of the transforming function must be taken in order to secure close results for large values of  $|\mu|$ .

It is difficult to set up a rigorous criterion as to the number of  $B'_s$  necessary to define adequately the transforming function, but the following considerations are of value in this connection.

Suppose that  $f(x)$  may be adequately represented in the interval  $a < x < b$  by  $m$  terms, i.e.

$$f(x) = A_0 + A_1 x + A_2 x^2 + \dots + A_m x^m,$$

and that  $m$  is large enough so that the first  $m$  terms of the expansion of the normal function give an adequate representation of it. This is clearly possible since the expansions with which we are dealing converge uniformly in the open interval. Then the first  $m$   $B'_s$  may be determined so that the first  $m$  terms of  $f(x)dx$  transformed by the transforming function determined by the  $m$   $B'_s$  are identical with the first  $m$  terms in the expansion of the normal function. In addition there will remain certain terms which may cause a serious discrepancy. For  $f(x)dx$  becomes

$$A_0 (B_1 + 2 B_2 u + \dots) + A_1 (B_1 u + B_2 u^2 + \dots) (B_1 + 2 B_2 u + \dots) + \dots + A_m (B_1 u + B_2 u^2 + \dots)^m (B_1 + 2 B_2 u + \dots).$$

Let us assume that all  $B'_i = 0$ ,  $i > m$ , and investigate the terms in  $u$  of degree higher than  $m$ .

Since the first terms of  $f(x)dx$  transformed contribute few terms involving  $u^n$ ,  $n > m$ , and the higher order terms have small coefficients, it is to be expected that if  $m$   $B'_s$  are used a good result will be obtained, at least for moderate values of  $u$ .

Some skewed distributions that differ considerably from normal may yield a rapidly converging sequence of  $B'_s$ , that is in case there is a natural relation of this kind existing between the non-normal and normal distributions.

The main reason for investigating the possibility of an easily determined transforming function that will transform a non-

normal distribution into a normal distribution is the fact that the distributions in random samples of estimates of the parameters of the non-normal distribution can be expressed in terms of the transformation and the sampling distributions of the parameters of the normal distribution into which the non-normal distribution is transformed. This proposition is developed in Part II.

## PART II

### DISTRIBUTION OF THE ESTIMATES OF THE PARAMETERS OF NON-NORMAL DISTRIBUTIONS

Suppose that a variable  $x$  is distributed as  $f(x)dx$  where  $f(x)$  is such that it can be transformed into a normal distribution by means of a quadratic transformation,

$$(1) \quad x = ay + by^2.$$

Then  $f(x)dx$  becomes

$$(2) \quad f(ay + by^2)(a + 2by)dy,$$

where  $y$  is normally distributed.

The total of a sample of  $n$   $x$ 's drawn at random from  $f(x)$  is

$$(3) \quad (x_1 + x_2 + x_3 + \dots + x_n),$$

which by virtue of (1) becomes

$$(4) \quad a(y_1 + y_2 + \dots + y_n) + b(y_1^2 + y_2^2 + \dots + y_n^2).$$

The coefficient of  $a$  in (4) is an estimate of the total of a sample of  $n$  of a normally distributed variable and the coefficient of  $b$  is an estimate of the second moment about a fixed point of a normally distributed variable which can be written as an estimate of the standard deviation squared plus the estimate of the mean squared. Thus (4) can be written as

$$n \cdot a \cdot \bar{m} + n \cdot b \cdot (\bar{\sigma}^2 + \bar{m}^2),$$

where the bar over  $m$  and  $\sigma$  denotes estimates of these parameters by means of samples. The distributions of  $\bar{m}$  and  $\bar{\sigma}$  are known and are independent. If the mean of distribution (2) is taken to be zero, then  $\bar{m}$  is distributed as proportional to  $e^{-\frac{n \cdot \bar{m}^2}{2}}$  and  $y = a\bar{m} + b\bar{m}^2$  is distributed as proportional to

$$(5) \quad \frac{e^{-\frac{1}{2} \cdot \frac{n(a^2 + 2by + a\sqrt{a^2 + 4by})}{2b^2}}}{\sqrt{a^2 + 4by}}, \quad -\frac{a}{2b} \leq y \leq \infty.$$

The distribution of  $b\bar{\sigma}^2$  is, except for a constant factor,

$$(6) \quad z^{\frac{n-3}{2}} e^{-\frac{nz}{2}}, \quad 0 \leq z \leq \infty$$

if  $n \geq 2$ .

If two variables,  $x$  and  $y$ , are distributed as  $f(x, y) dx dy$ , then the probability that a value of  $\varphi(x, y) = v$  is in  $dv$  is given as the surface area of the cylinder  $\varphi(x, y) = v$  between  $z = f(x, y)$  and  $z = 0$  times  $dv$ .<sup>3</sup>

In this case the probability function of  $v = y + z$  is proportional to

$$(7) \quad \int_{-\frac{a^2}{4b}}^v \frac{e^{-\frac{1}{2} \cdot \frac{n(a^2 + 2by + a\sqrt{a^2 + 4by})}{2b^2}}}{\sqrt{a^2 + 4by}} \cdot (v - y) \cdot e^{-\frac{n-3}{2} - \frac{n(v-y)}{2b}} dy.$$

Put  $y = ax + bx^2$  and (7) becomes

$$(8) \quad e^{-\frac{n}{2b}v} \int_{\frac{-a - \sqrt{a^2 + 4bv}}{2b}}^{\frac{-a + \sqrt{a^2 + 4bv}}{2b}} e^{\frac{na}{b}x} \cdot [v - ax - bx^2]^{\frac{n-3}{2}} dx, \quad -\frac{a^2}{4b} \leq v \leq \infty.$$

<sup>3</sup> Baker, G. A.—Random Sampling from Non-Homogeneous Populations, *Metron*, Vol. VIII, No. 3, Feb. 1930.

If  $f(x)$  can be transformed into a normal distribution by means of a cubic transformation.

$$(9) \quad x = ay + by^2 + cy^3$$

then (3) becomes

$$(10) \quad a(y_1 + y_2 + \dots + y_n) + b(y_1^2 + y_2^2 + \dots + y_n^2) + c(y_1^3 + y_2^3 + \dots + y_n^3)$$

which can be written as  $n(a\bar{m} + b\bar{m}^2 + c\bar{m}^3 + 3c\bar{\sigma}^2\bar{m} + b\bar{\sigma}^2)$ .

Hence the means of samples of  $n$  are distributed as proportional to

$$(11) \quad \int_{\gamma}^{\beta} \left[ \frac{\nu - ax - bx^2 - cx^3}{3cx + b} \right]^{\frac{n-3}{2}} e^{-\frac{nx^2}{2} - \frac{n}{2} \left[ \frac{\nu - ax - bx^2 - cx^3}{3cx + b} \right]} dx,$$

where  $\gamma/\beta$  represents the interval or intervals for which  $\nu - ax - bx^2 - cx^3$  and  $3cx + b$  have the same signs and  $\nu$  varies from  $-\infty$  to  $+\infty$ .

Suppose that the given frequency distribution can be transformed into a normal distribution by the transformation

$$x = ay + by^2,$$

and consider the expression for the estimation of the standard deviation squared of the  $x$ -distribution from a sample of  $n$ ,

$$(12) \quad \frac{(x_1^2 + x_2^2 + \dots + x_n^2)}{n} - \left[ \frac{(x_1 + x_2 + \dots + x_n)}{n} \right]^2$$

which becomes

$$(13) \quad \frac{[(ay_1 + by_1^2) + \dots + (ay_n + by_n^2)]}{n} - \left[ \frac{-(ay_1 + by_1^2) + \dots + (ay_n + by_n^2)}{n} \right]^2$$

where  $y$  is normally distributed. In terms of the estimates of the mean and standard deviation of the  $y$ 's (13) can be written as

$$2b^2\bar{\sigma}^4 + a^2\bar{\sigma}^2 + 4ab\bar{\sigma}^2\bar{m} + 4b^2\bar{\sigma}^2\bar{m}^2.$$

Hence the estimates of the standard deviations of the original population will be distributed as proportional to

$$(14) \quad \int_{-\infty}^{\infty} e^{-\frac{n}{2} \left[ \frac{-(a^2 + 4abx) + \sqrt{(a^2 + 4abx + 4b^2x^2)^2 + 8b^2v}}{4b^2} \right]} \cdot \left[ \frac{-(a^2 + 4abx + 4b^2x^2) + \sqrt{(a^2 + 4abx + 4b^2x^2)^2 + 8b^2v}}{4b^2} \right]^{\frac{7-3}{2}} \\ \sqrt{1 + \left[ -\frac{(4ab + 8b^2x)}{2b^2} + \frac{(a^2 + 4abx + 4b^2x^2)(4ab + 8b^2x)}{4b^2\sqrt{(a^2 + 4abx + 4b^2x^2)^2 + 8b^2v}} \right]^2} dx,$$

where  $v$  varies from 0 to  $+\infty$ .

The distributions of the estimates of other parameters and the distributions of the estimates of the mean and standard deviation for different transformations can be expressed in terms of the distributions of the mean and standard deviation of the resulting transformed normal distribution but it is obvious that the process becomes complicated.

# INVARIANTS AND COVARIANTS OF CERTAIN FREQUENCY CURVES

By

RICHMOND T. ZOCH

*Introduction.* After the most convenient type of equation  $y = f(x, a, b, c, \dots)$  has been selected and the parameters  $a, b, c, \dots$ , in the selected equation have been determined so that for a given set of values  $x_i$  ( $i = 1, 2, \dots, n$ ), the computed values  $y_i$  ( $i = 1, 2, \dots, n$ ) agree as closely as possible or as closely as is consistent with the observed values  $Y_i$  ( $i = 1, 2, \dots, n$ ), it may be desirable to make one or more of the transformations: (1) move the origin, (2) use a different scale (unit of measure), (3) change the total frequency.

This paper discusses certain invariants and covariants of the above transformations which were noted in developing the general theory for the Pearson Curves of frequency.

1. *Change of Origin.* Instead of considering the diff. eq.,

$$(1) \quad \frac{dy}{dx} = \frac{y(x-P)}{b_2 x^2 + b_1 x + b_0}$$

which is the diff. eq. from which the Pearson curves are derived, we take the more general diff. eq.,

$$(2) \quad \frac{dy}{dx} = \frac{y(x-P)}{b_n x^n + b_{n-1} x^{n-1} + \dots + b_1 x + b_0}$$

Equation (1) is a special case of equation (2).

Make the following substitutions:

$$(3) \quad \left\{ \begin{array}{l} x = X + P, \quad b_n = B_n, \\ n P b_n + b_{n-1} = B_{n-1}, \\ \frac{n(n-1)}{2!} P^2 b_n + (n-1) P b_{n-1} + b_{n-2} = B_{n-2}, \\ \dots \dots \dots \\ P^n b_n + P^{n-1} b_{n-1} + \dots \dots + b_0 = B_0, \end{array} \right.$$

and on simplifying we obtain,

$$(4) \quad \frac{dy}{dx} = \frac{yX}{B_n X^n + B_{n-1} X^{n-1} + \dots + B_1 X + B_0} = \frac{yX}{F(X)}.$$

If we now write

$$(5) \quad X = x - P$$

we have:

$$(6) \quad \frac{dy}{dx} = \frac{y(x-P)}{B_n (x-P)^n + B_{n-1} (x-P)^{n-1} + \dots + B_1 (x-P) + B_0}.$$

The solutions of equations (4) and (6) can be written in the form

$$(7) \quad y = G(X) = G(x-P),$$

where  $P$  is the mode as will be observed from the diff. eq. In other words the frequency function is a function of  $(x-P)$  when it is written in the form of eq. (7). Therefore if we change the origin of  $x$  by writing  $x' = x - h$  all of the constants of the frequency curve will remain unchanged if at the same time  $P$  be subjected to the transformation  $P' = P - h$ .

2. *Change of Total Frequency.* Let  $C_0$  be the constant of integration when the area under the curve is unity and when the argument is  $X = x - P$ ;  $K_0$  the constant of integration when the argument is  $X = x - P$  for an arbitrary area under the curve; and  $N$  the total frequency. Now when the total frequency is changed the area under the curve is changed, hence from the above definitions

$$(8) \quad K_0 = N C_0.$$

Therefore if the total frequency be  $N$  and it is desired to write the equation of the frequency function for a total frequency of  $\bar{N}$  then write  $\bar{K}_0$  for  $K_0$  where

$$(9) \quad \bar{K}_0 = (K_0 N) + \bar{N}$$

and leave all of the remaining constants unchanged.

It should be emphasized that in leaving the remaining con-



stants unchanged we assume that the distribution of the new sample or the universe obeys the same law as the old sample. Occasionally one sees the statement in works on probability and statistics in connection with the Theory of Errors that as the number of observations is increased indefinitely, the arithmetic mean tends to the true value of a distribution. This statement is based upon the tacit assumption that an observation less than the true value (most probable) is as likely to occur as an observation greater than the true value. If we make this assumption we will always (if the number of observations be sufficiently large) ultimately obtain a symmetrical frequency curve (the  $A, M$ . coincides with the axis of symmetry) and this assumption contradicts the assumption that the distribution of the new sample obeys the same law as the old sample (except the old sample itself be symmetrically distributed).

3. *Change of Scale.* We are now ready to consider the behavior of the constants when the unit of measure is changed. Perhaps it is well to point out here that quite often it is desirable to change the unit from months to years, from feet to yards, from pounds to grams, etc. The behavior of the constants under a change of scale is not as easily arrived at as for the changes of the origin and total frequency.

The behavior of  $B_n$  where  $B_n$  is the coefficient of the highest power of  $X$  in  $F(X)$  of the differential equation,  $\frac{dy}{dX} = \frac{yX}{F(X)}$ , will first be obtained.

Elderton<sup>1</sup> uses moments to determine the constants of a frequency curve. Thorkelsson<sup>2</sup> and Fisher<sup>3</sup> have used Thiele's semi-

<sup>1</sup> W. Palin Elderton, "Frequency Curves and Correlation", Second Edition 1927, London.

<sup>2</sup> Thorkell Thorkelsson, "Frequency Curves Determined by Semi-Invariants" (Visindafelag Islendinga IX) Reykjavik Ríkisprentsmíðjan Gutenberg.—MCMXXXI.

<sup>3</sup> Arne Fisher, "Frequency Curves", Translated by E. A. Vigfusson, American Edition, 1922, The Macmillan Co.

invariants for this purpose. Semi-invariants have an advantage over moments in that the values of the higher semi-invariants do not change when the origin is changed. Moreover Fisher (pp. 12-16, loc. cit.) has pointed out how the semi-invariants behave when the unit is changed, viz:

$$\lambda_1(ax+c) = a \lambda_1(x) + c$$

$$\lambda_i(ax+c) = a^i \lambda_i(x) \text{ for } i > 1.$$

Referring to equation (2) let  $P_0$  be the value of  $P$  when the origin is at the arithmetic mean, and let  $t'_n, t'_{n-1}, \dots, t'_1$ , and  $t'_0$  be the values of  $t_n, t_{n-1}, \dots, t_1$ , and  $t_0$  when the origin is at the arithmetic mean. Now Thorkelsson (loc. cit.) has pointed out that when his method is used for computing the constants of the curve there will be only one equation involving  $P_0$  and only one equation involving  $t'_0$ . Moreover the coefficients of the  $(t')$ 's and the constant terms of the remaining equations will be of constant weight.

Below is an example of the equations obtained when Thorkelsson's method is used to compute the constants:

$$(10) \left\{ \begin{array}{l} -P_0 + t'_1 + 3\lambda_2 t'_3 = 0 \\ \lambda_2 + t'_0 + 3\lambda_2 t'_2 + 4\lambda_3 t'_3 = 0 \\ \lambda_3 + 2\lambda_2 t'_1 + 4\lambda_3 t'_2 + (5\lambda_4 + 12\lambda_2^2) t'_3 = 0 \\ \lambda_4 + 3\lambda_3 t'_1 + (5\lambda_4 + 6\lambda_2^2) t'_2 + (6\lambda_5 + 45\lambda_2\lambda_3) t'_3 = 0 \\ \lambda_5 + 4\lambda_4 t'_1 + (6\lambda_5 + 24\lambda_2\lambda_3) t'_2 + (7\lambda_6 + 72\lambda_2\lambda_4 + 54\lambda_3^2 + 24\lambda_2^3) t'_3 = 0 \end{array} \right.$$

Note that only the first of the above equations involves  $P_0$  and only the second involves  $\ell'_0$ .

Since the coefficients are of constant weight they are *invariants*<sup>4</sup> of index  $w$  where  $w$  is the weight of the coefficient when  $x$  is subjected to the transformation  $x' = ax + c$ .

Suppose that we now consider the general case where  $F(X)$  is of degree  $n$ . Hence, in general, equations (10) will consist of  $n+2$  equations in  $n+2$  unknowns; the unknowns being  $P_0, \ell'_0, \ell'_1, \dots, \ell'_n$ . Disregard the two equations which involve  $P_0$  and  $\ell'_0$  then there remain  $n$  equations in  $n$  unknowns. Observe that the weights of the coefficients of the  $\ell'_i$  form an A.P. whether taken by rows or by columns. Also the weights of the constant terms form the same A.P. as the columns.

We now state the

**Lemma:** If all of the elements of a determinant are covariants and the weights (indices) of the elements of every row form an A.P. and of every column form an A.P. then when the determinant is expanded every term is of constant weight (index).

**Proof:** Let the A.P. formed by the weights of the elements of the rows be  $w_{ni} = a_n + (i-1)\delta$   $\begin{cases} n = 1, 2, \dots, n \\ i = 1, 2, \dots, n. \end{cases}$

Then the weights of the elements can be displayed as follows:

$$\begin{array}{cccccc}
 a_1 & a_1 + \delta & a_1 + 2\delta & a_1 + 3\delta & \dots & a_1 + (n-1)\delta \\
 a_2 & a_2 + \delta & a_2 + 2\delta & a_2 + 3\delta & \dots & a_2 + (n-1)\delta \\
 a_3 & a_3 + \delta & a_3 + 2\delta & a_3 + 3\delta & \dots & a_3 + (n-1)\delta \\
 \dots & \dots & \dots & \dots & \dots & \dots \\
 a_n & a_n + \delta & a_n + 2\delta & a_n + 3\delta & \dots & a_n + (n-1)\delta
 \end{array}$$

(It should be emphasized that the above is not the determinant mentioned in the statement of the Lemma but the elements of the above array represent the weights of the elements of the determi-

<sup>4</sup>L. E. Dickson, "Modern Algebraic Theories", Benj. H. Sanborn & Co., 1926; Chicago. Chapter I.

nant). Now since by hypothesis the elements of every column have such weights that the weights form an A.P. then  $a_1, a_2, a_3, \dots, a_n$  must form an A.P. Let this A.P. be  $a_1 + (j-1)\bar{\delta}$ . Making use of this notation the weights of the elements of the determinant can be displayed as follows:

$$\begin{array}{cccc} a_1 & a_1 + \delta & \dots & a_1 + (n-1)\delta \\ a_1 + \bar{\delta} & a_1 + \delta + \bar{\delta} & \dots & a_1 + (n-1)\delta + \bar{\delta} \\ \dots & \dots & \dots & \dots \\ a_1 + (n-1)\bar{\delta} & a_1 + \delta + (n-1)\bar{\delta} & \dots & a_1 + (n-1)\delta + (n-1)\bar{\delta} \end{array}$$

Hence the weight of the element in the  $i^{th}$  row and the  $j^{th}$  column is  $a_1 + (i-1)\delta + (j-1)\bar{\delta}$ . Along the principal diagonal of the determinant  $i=j$ . Therefore when the determinant is expanded the weight of the term consisting of the elements of the principal diagonal is the sum of the A.P.  $w_i = a_1 + (i-1)(\delta + \bar{\delta})$  or

$$\sum_{i=1}^n w_i = \frac{n}{2} \cdot [2a_1 + (n-1)(\delta + \bar{\delta})] = W.$$

Every term in the expansion is of weight  $W$  because each term consists of one element from each row and one element from each column and hence the weight is equal to the sum of two series, each being an A.P., plus the weight of the term in the upper left corner.

**THEOREM:** If all of the coefficients and the "constant" terms of a system of  $n$  linear equations in  $n$  unknowns are covariants of such respective weights (indices) that the weights (indices) of the elements of every row of the matrix of the system of equations form an A.P. and of the elements of every column of the augmented matrix form an A.P. then the solutions are covariants whose weights (indices) form an A.P. whose common difference is of the same magnitude but of opposite sign to the common difference in the A.P. of the weights (indices) of the elements of the rows.

Proof: By Cramer's rule the solutions are

$$z_i = \frac{D_i}{\Delta} \quad \text{where } \Delta = \begin{vmatrix} K_{11} & \dots & K_{1n} \\ \vdots & & \vdots \\ K_{n1} & \dots & K_{nn} \end{vmatrix} \quad \text{and where}$$

$D_i$  is the  $n$ -rowed determinant obtained from  $\Delta$  by replacing the elements of the  $i^{th}$  column by the "constant" terms of the system. Let the weight (index) of the element in the  $i^{th}$  row and  $j^{th}$  column of  $\Delta$  be  $a_i + (i-1)\delta + (j-1)\bar{\delta}$ . Also let the A.P. formed by the weights (indices) of the elements of the  $n^{th}$  row of  $\Delta$  be  $w_{ni} = a_n + (i-1)\bar{\delta}$ , hence in particular the A.P. of the weights (indices) of the elements of the first row are  $a_1 + (i-1)\bar{\delta}$ . Further let the A.P. formed by the elements of the column of constant terms of the augmented matrix be  $w_{ci} = a_c + (i-1)\bar{\delta}$ . By the lemma just established we see that when  $\Delta$  is expanded all of the terms of the expansion will be of the same weight (index)  $W$ . Hence  $\Delta$  is of weight (index)  $W$ . Also since the A.P. of the weights (indices) of the column of constant terms is  $w_{ci} = a_c + (i-1)\bar{\delta}$  then the weight (index) of each term in the expansion of  $D_i$  will be  $-[a_i + (i-1)\bar{\delta}] + a_c$  different from the weight (index) of each term in the expansion of  $W$ . Hence the weight (index) of  $D_i$  is  $W - [a_i + (i-1)\bar{\delta}] + a_c$ . Therefore the weight (index) of  $z_i$  is  $W - [a_i + (i-1)\bar{\delta}] + a_c - W = a_c - a_i + (i-1)(-\bar{\delta})$ , and the theorem is established.

Applying the above Theorem and observing that  $\delta = 1$  in equations (10) we obtain the result:

$$\text{weight of } b'_n = 3 - 2 + (n-1)(-1) = 2 - n.$$

Since  $B_n = b'_n$  we have the result that when  $x$  is subjected to the transformation  $x' = ax + c$  then  $B'_n = a^{2-n} B_n$ . Or in other words  $B_n$  is an invariant of index  $2-n$  under the transformation  $x' = ax + c$ .

Now we turn to the consideration of  $P_0$ . Here we have

$n+1$  equations in  $n+1$  unknowns and the augmented matrix has elements of the following weights (\* means that an element is lacking):

$P_0$	0	*	2	3	...	$n-1$	*
*	2	3	4	5	...	$n+1$	3
*	3	4	5	6	...	$n+2$	4
*	4	5	6	7	...	$n+3$	5
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
*	$(n+1)$	$(n+2)$	$(n+3)$	$(n+4)$	...	$2n$	$(n+2)$

Now  $P_0$  is the quotient of two determinants formed from the above matrix and if these two determinants be expanded in terms of the minors of the first column we see that the weight of  $(W+1) - W = 1$ . That is  $P_0$  is of weight 1 regardless of the degree of  $F(X)$ . Therefore  $P_0$  is an invariant of index 1.

Next considering  $\ell'_0$  the augmented matrix has the same elements as for  $P_0$  except that the first row is now:

$\ell'_0$  \* 2 3 4 ...  $n-1$   $\frac{2}{\ell'_0}$   
 Following the same procedure we see that the weight of  $\ell'_0 = (W+2) - W = 2$ . Therefore  $\ell'_0$  is an invariant of index 2.

We can look upon equations (3) as a transformation. We can reverse this transformation by solving for the  $\ell_2$  in terms of the  $B_i$ . Also, by moving the origin to the A.M. equations (3) may be written:

$$(11) \left\{ \begin{array}{l} B_n = \ell'_n \\ B_{n-1} = n C_n P_0 \ell'_n + \ell'_{n-1} \\ B_{n-2} = n C_{n-1} P_0^2 \ell'_n + n C_{n-1} P_0 \ell'_{n-1} + \ell'_{n-2} \\ \vdots \\ B_{n-r} = n C_{n-r+1} P_0^r \ell'_n + n C_{n-r+1} P_0^{r-1} \ell'_{n-1} + \dots + \ell'_{n-r} \end{array} \right.$$

In equations (11)  $P_o, b'_n, b'_{n-1}, \dots, b'_o$  are the values of  $P_o, b_n, b_{n-1}, \dots, b_o$  when the origin is at the A.M.

Note that the right hand numbers of equations (11) are isobaric and that  $B_o$  is of weight 2;  $B_1$  of weight 1;  $B_2$  of weight 0; and in general  $B_i$  is of weight  $2-i$ .

Now let  $g_o = \frac{b'_o}{b'_n}, g_1 = \frac{b'_1}{b'_n}$ , in general  $g_i = \frac{b'_i}{b'_n}$ ; hence

$g_n = 1$ . Therefore when the  $g$ 's are computed we note that  $g_o$  is of weight  $n$ ;  $g_1$  is of weight  $n-1$  and in general  $g_i$  is of weight  $(2-i) - (2-n) = n-i$ . Since  $g_o$  is the product of all the roots,  $g_1$  the sum of the products taken  $(n-1)$  at a time and so on and  $g_{n-1}$  is the sum of all the roots (due consideration being taken of the signs) it follows that all the roots of  $F(X)$  are invariants of index 1 under the transformation  $x' = ax + c$ .

Now if equation (4) be solved in the form of equation (7) then it can be seen by actual substitution of the indices of  $B$  and the roots of  $F(X)$  which are involved in the constants that the exponents  $\kappa$  and  $\epsilon$  of factors of the form  $(1 - \frac{x-P}{\lambda_i})^\kappa$  and

$$\frac{e^{\epsilon_1 \epsilon_2 \arctan \frac{x-P+\lambda_3}{\lambda_o}}}{\left[1 + \left(\frac{x-P+\lambda_3}{\lambda_o}\right)^2\right]^{\epsilon_1/2}} \quad \text{are invariants of index zero. The}$$

factor  $(1 - \frac{x-P}{\lambda_i})^\kappa$  occurs for a real root  $\lambda_i$  of  $F(X)$  and the factor

$$\frac{e^{\epsilon_1 \epsilon_2 \arctan \frac{x-P+\lambda_3}{\lambda_o}}}{\left[1 + \left(\frac{x-P+\lambda_3}{\lambda_o}\right)^2\right]^{\epsilon_1/2}} \quad \text{occurs for a pair of conjugate com-}$$

plex roots of  $F(X)$ . The fact that the exponents  $\kappa$  are invariants of index zero will be generalized for the case where complex roots do not occur and where no real root is repeated.

If complex roots do not occur the differential equation

$$\frac{dy}{dx} = \frac{y X}{F(X)} \text{ can be written } \frac{dy}{y} = \frac{X dX}{F(X)} = \frac{1}{B_n} \left[ \frac{m_1}{X+\lambda_1} + \dots + \frac{m_n}{X+\lambda_n} \right] dX$$

where in separating  $\frac{X}{F(X)}$  into partial fractions and equating coefficients of like powers of  $X$  we obtain  $n$  equations in  $n$  unknowns and since the roots are all of weight 1 the weights of the augmented matrix will be (the unknowns of the system are the  $m_i$ ):

$$\begin{array}{ccccccccccc} n-1 & n-1 & n-1 & . & . & . & . & . & n-1 & * \\ n-2 & n-2 & n-2 & . & . & . & . & . & n-2 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & . & . & . & . & . & 0 & * \end{array}$$

Applying the Lemma we see that the  $m_i$  are all of the same weight (since  $\delta = 0$ ). Expanding the determinant in the numerator by minors we see that the  $m_i$  are of weight  $n-2$ . Since  $B_n$  is of weight  $2-n$ , we have  $\frac{m_i}{B_n} = \kappa_i$  is of weight zero. Therefore the  $\kappa_i$  are invariants of index zero under the transformation

$$x' = ax + c.$$

We have now considered all of the constants of the curves except the constant of integration. Let the solutions be written in the form:  $y = C_0 G(X)$ .

Now it is possible to write  $G(X)$  in such a form that  $G(X)$  is a *covariant* of index zero under the transformation  $X' = aX$ . In the case of real roots this is accomplished by dividing both the numerator and the denominator of each partial fraction by the root involved in the fraction before the integration is performed. Partial fractions which involve complex roots can be similarly treated. This is the way Pearson actually treated his Types I, II, III, IV, and VII curves although he did not deal with his Types V and VI curves in this manner.

After we have our solution in the form which makes  $G(X)$  a covariant of index zero then if we write  $X'$  for  $aX$  the total



frequency between  $nX'$  and  $(n+1)X'$  will be the same as the total frequency between  $naX$  and  $(n+1)aX$ . Therefore  $y$  is a co-variant of index  $(-1)$ . Hence  $C_0$  is an invariant of index  $(-1)$ .

An example will now be given. Take the equation to which Elderton (loc. cit.) fits a Type I curve on pages 54-59. He has used a unit of 5 years. Suppose we wish to change to a unit of 1 year. Then the constants  $a_1$  and  $a_2$  being the roots of  $F(X)$  are invariants of index 1 and are each multiplied by 5 and become 9.98190 and 67.63640 respectively. Since  $m_1$  and  $m_2$  are invariants of index zero they remain unchanged and are as he gives them viz. .409833 and 2.776978. The constant of integration being an invariant of index  $-1$  it is divided by 5 and becomes 29.892. The equation with a unit of 1 year becomes (See top of page 58):

$$y = 28.892 \left\{ 1 + \frac{x'}{9.98190} \right\}^{.409833} \cdot \left\{ 1 - \frac{x'}{67.63640} \right\}^{2.776978}$$

Suppose that now we wish to move the origin to age 26.75942. Then the above equation becomes:

$$y = 28.892 \left\{ 1 + \frac{x'' - 26.75942}{9.98190} \right\}^{.409833} \cdot \left\{ 1 - \frac{x'' - 26.75942}{67.63640} \right\}^{2.776978}$$

Finally suppose we wish to change to a total frequency of 2000 instead of 1000 as in the given sample. Then the equation becomes:

$$y'' = 59.784 \left\{ 1 + \frac{x'' - 26.75942}{9.98190} \right\}^{.409833} \cdot \left\{ 1 - \frac{x'' - 26.75942}{67.63640} \right\}^{2.776978}$$

4. *Conclusion: Benefits of this Information.* If the diff. eq. (2) be written in the form (4) by means of the transformation (3) then the integration is more easily accomplished. That is to say: in general the solution in the form of eq. (7) is more readily obtained from (4) than some equivalent form of solution would be from (2). Thus a solution in the form of eq. (7) is not only

more easily obtained but also lends itself readily to a change of origin.

Each type of Pearson's Curves may be written in a number of ways. The numerical example given above shows the convenience and advantage of writing a solution so that the origin is at the mode,  $G(X)$  is a covariant of index zero,  $y$  a covariant of index  $(-1)$  and the constant of integration an invariant of index  $(-1)$ .

Regardless of what form is selected for writing a solution the solution will be a covariant and the constants will be invariants, but not necessarily of the indices mentioned above. A knowledge of these invariants will save much labor if it is desired to make a change in the unit of measure.

Similar laws of transformation can be worked out for (1) solutions of the diff. eq.  $\frac{dy}{dx} = \frac{y f(x)}{F(x)}$  where both  $f(x)$  and  $F(x)$  are integral rational functions of  $x$  and (2) for the Gram-Charlier Types A and B series. In the first case we obtain the same result as outlined above for the simpler diff. eq.  $\frac{dy}{dx} = \frac{y x}{F(x)}$ ; that is the solution may be written in the form  $y = C_0 \cdot G(X)$  where  $G(X)$  is a covariant of index zero,  $y$  a covariant of index  $(-1)$  and  $C_0$  is an invariant of index  $(-1)$ . In the case of the Type A series the coefficient of each term is an invariant of index zero.

George Washington University.

# QUADRATURE OF THE NORMAL CURVE

By

E. R. ENLOW

There are three formulas for the calculation of areas under the normal probability curve, only two of which seem to be generally recognized in American statistical circles. Herewith is presented an outline of the mathematical development of these three formulas and a determination of the bounds of practical utility of each.

The well-known equation for the normal curve,

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

may be expanded into the series

$$y = \frac{N}{\sigma\sqrt{2\pi}} \left[ 1 - \left(\frac{x}{\sigma\sqrt{2}}\right)^2 + \frac{1}{12} \left(\frac{x}{\sigma\sqrt{2\pi}}\right)^4 - \frac{1}{15} \left(\frac{x}{\sigma\sqrt{2\pi}}\right)^6 + \dots \right] \quad (\text{Ref. 3})$$

by means of Maclaurin's Theorem. (See any good calculus text-book.) (7) This expansion is readily accomplished by making the substitution

$$t = \frac{x}{\sigma\sqrt{2}}$$

so that

$$e^{-\frac{x^2}{2\sigma^2}} = e^{-t^2}$$

and

$$f(t) = e^{-t^2}$$

The process of successive differentiation is quite lengthy, since every other term differentiated becomes zero and therefore 2n terms in the Maclaurin series are required to produce n terms in the new series. After the expansion has been carried to five or six terms, a regular law of formation becomes evident from inspection of the new series

$$e^{-t^2} = 1 - t^2 + \frac{1}{2} t^4 - \frac{1}{6} t^6 + \frac{1}{24} t^8 - \dots$$

viz.:

$$n\text{th term} = \frac{1}{L^{n-1}} t^{2n-2}$$

After making the reversion  $t = \frac{x}{\sigma\sqrt{2}}$

and substituting the value of  $e^{-\frac{x^2}{2\sigma^2}}$  in the original equation for the normal curve, we have

$$y = \frac{N}{\sigma\sqrt{2\pi}} \left[ 1 - \left( \frac{x}{\sigma\sqrt{2}} \right)^2 + \frac{1}{L^2} \left( \frac{x}{\sigma\sqrt{2}} \right)^4 - \frac{1}{L^3} \left( \frac{x}{\sigma\sqrt{2}} \right)^6 + \dots \right] \quad (\text{Ref. 1, 7})$$

as previously indicated. This series is uniformly convergent and may therefore be integrated term by term.

The area under all or any portion of the normal curve is calculated from the integral of the equation for the curve:

$$\int y = \frac{N}{\sigma\sqrt{2\pi}} \int e^{-\frac{x^2}{2\sigma^2}} dx.$$

To simplify the procedure, let  $x = \sigma\sqrt{2} \cdot t$ .

Then  $dx = \sigma\sqrt{2} \cdot dt$

$$\begin{aligned} \text{and } \int y &= \frac{N}{\sigma\sqrt{2\pi}} \int e^{-t^2} \cdot \sigma\sqrt{2} \cdot dt = \frac{N}{\sqrt{\pi}} \int e^{-t^2} dt \\ &= \frac{1}{\sqrt{\pi}} \int e^{-t^2} dt \quad (\text{when } N=1). \end{aligned}$$

The value of the definite integral representing the area between the ordinate at the mean and the general ordinate whose abscissa is  $t$  is

$$\int_0^t y = \frac{1}{\sqrt{\pi}} \int_0^t e^{-t^2} dt.$$

Then, integrating the expanded series for  $e^{-t^2}$  (above) term by term, we have:

$$\begin{aligned} \frac{1}{\sqrt{\pi}} \int_0^t \left[ 1 - t^2 + \frac{t^4}{L^2} - \frac{t^6}{L^3} + \frac{t^8}{L^4} - \dots \right] dt = \\ \frac{t}{\sqrt{\pi}} \left[ 1 - \frac{t^2}{2} + \frac{t^4}{4L^2} - \frac{t^6}{6L^3} + \frac{t^8}{8L^4} - \frac{t^{10}}{10L^5} + \frac{t^{12}}{12L^6} - \dots \right]. \end{aligned}$$

Substituting in this series the value of  $t = \frac{x}{\sigma\sqrt{2}}$  and keeping the expression  $\frac{x}{\sigma}$  separate, we obtain:

$$\int_0^{\frac{x}{\sigma}} y = \frac{x}{\sigma\sqrt{2\pi}} \left\{ 1 - \frac{\left(\frac{x}{\sigma}\right)^2}{2 \cdot 3} + \frac{\left(\frac{x}{\sigma}\right)^4}{2 \cdot 5 \cdot 2^2} - \frac{\left(\frac{x}{\sigma}\right)^6}{2 \cdot 3 \cdot 7 \cdot 2^3} + \frac{\left(\frac{x}{\sigma}\right)^8}{2 \cdot 4 \cdot 9 \cdot 2^4} - \dots \right\}.$$

This series may be extended indefinitely, since the general term is seen to be

$$\frac{\left(\frac{x}{\sigma}\right)^{2n-2}}{2^{n-1} (2n-1) 2^{n-1}}.$$

It will be referred to hereinafter as Series A.

A published statement that Series A is divergent when  $t > 1$  is erroneous (7). It is always convergent, regardless of the value of the deviate, but converges very slowly when  $t$  is not small, in which case it is better to use another series obtained by integrating by parts. (1, 7)

We may write

$$\begin{aligned} \int e^{-t^2} dt &= \int \underbrace{[-\frac{1}{2t}]}_{[u]} \underbrace{[-2t e^{-t^2} dt]}_{[dv]} \\ &= \int \underbrace{[-\frac{1}{2t}]}_{[u]} \underbrace{[d(e^{-t^2})]}_{[dv]} \end{aligned}$$

Then

$$\int e^{-t^2} dt = -\frac{1}{2t} (e^{-t^2}) - \frac{1}{2} \int \frac{e^{-t^2}}{t^2} dt$$

$$(\text{Formula}) \quad \boxed{\int u dv} = \boxed{u \cdot v} - \boxed{\int v du}$$

Integrating the integral expression in the last term above, i. e.,

$$-\frac{1}{2} \int \frac{e^{-t^2}}{t^2} dt$$

in like manner (by parts), another term appears, and the equation above becomes

$$\int e^{-t^2} dt = -\frac{e^{-t^2}}{2t} + \frac{e^{-t^2}}{4t^3} + \frac{3}{4} \int \frac{e^{-t^2}}{t^4} dt.$$

Continuing this process by breaking up the integral on the right into another term in the series plus a new integral, repeatedly, produces the infinite series:

$$\begin{aligned} \int e^{-t^2} dt &= -\frac{e^{-t^2}}{2t} + \frac{e^{-t^2}}{4t^3} - \frac{3e^{-t^2}}{8t^5} + \frac{3 \cdot 5 e^{-t^2}}{16t^7} - \frac{3 \cdot 5 \cdot 7 e^{-t^2}}{32t^9} + \dots \\ &= -\frac{e^{-t^2}}{2t} \left[ 1 - \frac{1}{2t^2} + \frac{1 \cdot 3}{(2t^2)^2} - \frac{1 \cdot 3 \cdot 5}{(2t^2)^3} + \frac{1 \cdot 3 \cdot 5 \cdot 7}{(2t^2)^4} - \dots \right]. \end{aligned}$$

Now 
$$\int_0^t e^{-t^2} dt = \int_0^\infty e^{-t^2} dt - \int_t^\infty e^{-t^2} dt.$$

$$\boxed{\text{part}} = \boxed{\text{whole}} - \boxed{\text{part}}$$

But

$$\int_0^\infty e^{-t^2} dt = \frac{\sqrt{\pi}}{2} \quad * \quad (\text{Ref. 4})$$

Therefore

$$\int_0^t e^{-t^2} dt = \frac{\sqrt{\pi}}{2} - \int_t^\infty e^{-t^2} dt.$$

And since the value of the definite integral

$$\int_t^\infty e^{-t^2} dt = \left[ -\frac{e^{-t^2}}{2t} \left\{ 1 - \frac{1}{2t^2} + \frac{1 \cdot 3}{(2t^2)^2} - \frac{1 \cdot 3 \cdot 5}{(2t^2)^3} + \dots \right\} \right]_t^\infty.$$

Then

$$\int_0^t e^{-t^2} dt = \frac{\sqrt{\pi}}{2} - \frac{e^{-t^2}}{2t} \left\{ 1 - \frac{1}{2t^2} + \frac{1 \cdot 3}{(2t^2)^2} - \dots \right\},$$

and, since

$$\frac{N}{\sigma\sqrt{2\pi}} \int e^{-\frac{x^2}{2\sigma^2}} dx = \frac{1}{\sqrt{\pi}} \int e^{-t^2} dt,$$

$$(\text{when } N=1 \text{ and } t = \frac{x}{\sigma\sqrt{2}})$$

\* Good proof in The Encyclopedia Britannica, American Edition, 1896, in article on Infinitesimal Calculus. Also (2).

$$\frac{1}{\sqrt{\pi}} \int_0^t e^{-t^2} dt = \frac{1}{2} \left[ 1 - \frac{e^{-t^2}}{t\sqrt{\pi}} \left\{ 1 - \frac{1}{2t^2} + \frac{3}{(2t^2)^2} - \frac{3 \cdot 5}{(2t^2)^3} + \frac{3 \cdot 5 \cdot 7}{(2t^2)^4} - \dots \right\} \right].$$

Substituting  $\frac{x}{\sigma\sqrt{2}}$  for  $t$  and keeping  $\frac{x}{\sigma}$  separate, this series becomes:

$$\frac{1}{2} - \frac{e^{-\frac{x^2}{2\sigma^2}}}{\frac{x}{\sigma}\sqrt{2\pi}} \left\{ 1 - \frac{1}{\left(\frac{x}{\sigma}\right)^2} + \frac{3}{\left(\frac{x}{\sigma}\right)^4} - \frac{3 \cdot 5}{\left(\frac{x}{\sigma}\right)^6} + \frac{3 \cdot 5 \cdot 7}{\left(\frac{x}{\sigma}\right)^8} - \dots \right\}$$

This series will be referred to as Series B. It is asymptotic or semi-convergent, (5) (8), a type of series which is frequently obtained by integrating by parts (6). Series B is divergent for values of  $\frac{x}{\sigma}$  below unity. Weld (7) states that this series converges rapidly when  $\frac{x}{\sigma} > \sqrt{2}$ , but does not mention its peculiar asymptotic nature whereby it converges until a minimum term is reached and then diverges. As Townsend (6) indicates, the best approximation of the sum of an asymptotic series is obtained if the series is terminated with the term having the smallest absolute value. This minimum term is the second term for  $\frac{x}{\sigma} = \sqrt{2}$  and while the error due to dropping the succeeding terms is less than the last term retained, still this second term has too large a value to permit any very accurate calculation of the area (as will be shown later). However, the accuracy increases as  $\frac{x}{\sigma}$  takes on larger values, since it then takes longer for convergence to the minimum term and this minimum term also grows smaller.

Brunt (1) advocates the use of another series, developed by Schlömilch, which he states is better when  $\frac{x}{\sigma}$  is large. This

Schlömilch series, hereinafter referred to as Series S, is as follows, in terms of  $\frac{x}{\sigma}$  :

$$\text{Area} \int_0^{\frac{x}{\sigma}} = \frac{1}{2} \left[ 1 - \frac{2e^{-\frac{1}{2}(\frac{x}{\sigma})^2}}{\frac{x}{\sigma} \sqrt{2\pi}} \left\{ 1 - \frac{1}{(\frac{x}{\sigma})^2 + 2} + \frac{1}{\{(\frac{x}{\sigma})^2 + 2\} \{(\frac{x}{\sigma})^2 + 4\}} - \right. \right. \\ \left. \frac{5}{\{(\frac{x}{\sigma})^2 + 2\} \{(\frac{x}{\sigma})^2 + 4\} \{(\frac{x}{\sigma})^2 + 6\}} + \frac{9}{\{(\frac{x}{\sigma})^2 + 2\} \{(\frac{x}{\sigma})^2 + 4\} \{(\frac{x}{\sigma})^2 + 6\} \{(\frac{x}{\sigma})^2 + 8\}} \right. \\ \left. - \frac{129}{\{(\frac{x}{\sigma})^2 + 2\} \dots \{(\frac{x}{\sigma})^2 + 10\}} + \frac{315^*}{\{(\frac{x}{\sigma})^2 + 2\} \dots \{(\frac{x}{\sigma})^2 + 12\}} \dots \right]$$

Series S is readily developed from series B by transformation of successive terms in the B series to terms with the characteristic Schlömilch denominator. This is more easily accomplished when series B is in the " $t$ " ( $= \frac{x}{\sigma\sqrt{2}}$ ) form.

To determine the limits of practical utility of each of these three series, actual calculations of areas were made at appropriate  $\frac{x}{\sigma}$  intervals, with results shown in Table 1 and Figure 1. All calculations were made "by hand" and carried to 10 or more decimal places.

The three series (formulas) were used in the following forms:

SERIES A:—

$$\text{Area} \int_0^{\frac{x}{\sigma}} = \frac{x}{\sigma} (398\,942\,280\,3) \left[ 1 - \frac{(\frac{x}{\sigma})^2}{6} + \frac{(\frac{x}{\sigma})^4}{40} - \frac{(\frac{x}{\sigma})^6}{336} \right. \\ \left. + \frac{(\frac{x}{\sigma})^8}{3456} - \frac{(\frac{x}{\sigma})^{10}}{42240} + \frac{(\frac{x}{\sigma})^{12}}{599040} - \frac{(\frac{x}{\sigma})^{14}}{9676800} + \frac{(\frac{x}{\sigma})^{16}}{175472640} \right. \\ \left. - \frac{(\frac{x}{\sigma})^{18}}{3530096640} + \frac{(\frac{x}{\sigma})^{20}}{78033715200} - \frac{(\frac{x}{\sigma})^{22}}{1880240947200} + \dots \right]$$

\* This term not given by Brunt (1), but calculated by present writer. Last term practicable to use, since next term also has plus sign.



ERIES B:—

$$\text{Area} \int_0^{\frac{x}{\sigma}} = \frac{1}{2} - \log^{-1} \left[ 0 - \left\{ \log \frac{x}{\sigma} + \frac{\left(\frac{x}{\sigma}\right)^2}{2} (.434\ 294\ 481\ 9) + .399\ 089\ 934\ 2 \right\} \cdot \left[ 1 - \frac{1}{\left(\frac{x}{\sigma}\right)^2} + \frac{3}{\left(\frac{x}{\sigma}\right)^4} - \frac{15}{\left(\frac{x}{\sigma}\right)^6} + \frac{105}{\left(\frac{x}{\sigma}\right)^8} - \frac{945}{\left(\frac{x}{\sigma}\right)^{10}} + \frac{10\ 395}{\left(\frac{x}{\sigma}\right)^{12}} - \frac{135\ 135}{\left(\frac{x}{\sigma}\right)^{14}} + \frac{02\ 702\ 5}{\left(\frac{x}{\sigma}\right)^{16}} - \frac{344\ 594\ 25}{\left(\frac{x}{\sigma}\right)^{18}} + \frac{654\ 729\ 075}{\left(\frac{x}{\sigma}\right)^{20}} - \frac{137\ 493\ 105\ 75}{\left(\frac{x}{\sigma}\right)^{22}} + \dots \right] \right]$$

ERIES S:—

$$\text{Area} \int_0^{\frac{x}{\sigma}} = \frac{1}{2} - \log^{-1} \left[ 0 - \left\{ \log \frac{x}{\sigma} + \frac{\left(\frac{x}{\sigma}\right)^2}{2} (.434\ 294\ 481\ 9) + .399\ 089\ 934\ 2 \right\} \cdot \left[ 1 - \frac{1}{\left(\frac{x}{\sigma}\right)^2+2} + \frac{1}{\left\{ \left(\frac{x}{\sigma}\right)^2+2 \right\} \left\{ \left(\frac{x}{\sigma}\right)^2+4 \right\}} - \frac{5}{\left\{ \left(\frac{x}{\sigma}\right)^2+2 \right\} \left\{ \left(\frac{x}{\sigma}\right)^2+4 \right\} \left\{ \left(\frac{x}{\sigma}\right)^2+6 \right\}} + \frac{9}{\left\{ \left(\frac{x}{\sigma}\right)^2+2 \right\} \dots \left\{ \left(\frac{x}{\sigma}\right)^2+8 \right\}} - \frac{129}{\left\{ \left(\frac{x}{\sigma}\right)^2+2 \right\} \dots \left\{ \left(\frac{x}{\sigma}\right)^2+10 \right\}} + \frac{315}{\left\{ \left(\frac{x}{\sigma}\right)^2+2 \right\} \dots \left\{ \left(\frac{x}{\sigma}\right)^2+12 \right\}} - \dots \right] \right]$$

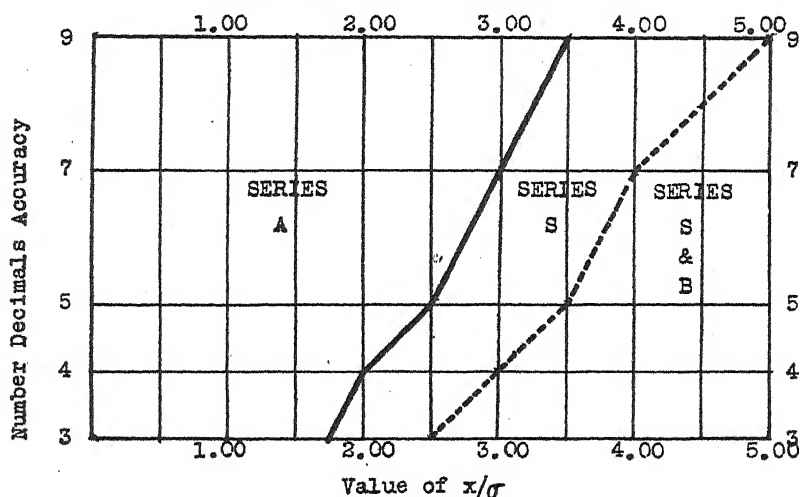
Table 1 shows the numbers of terms required in each series to give areas accurate to 3, 4, 5, 7 and 9 decimal places, respectively, for values of  $\frac{x}{\sigma}$  ranging from .25 to 5.00. Calculations were checked by Sheppard's Tables (4) and the accuracy was determined on the principle that the error is less than the last term retained or the first term dropped\*. Where x is used it indicates that the series cannot give the accuracy indicated.

The graph, Figure 1, shows the approximate domain of

\* This does not hold strictly true of Series S, since it is a modification of the true Series B.

utility of each series under conditions of accuracy ranging from 3 to 9 decimal places. Note that Series S covers a wider range than does Series B, including the entire domain of Series B. Hence we may conclude that, while it is essential as a basis for the derivation of Series S, Series B may be discarded for area calculations. Moreover, Series S is not only more valuable than Series B because of its wider range of utility but also because its more rapid convergence gives a desired degree of accuracy with fewer terms than are necessary with Series B.

FIGURE 1  
DOMAINS OF PRACTICAL UTILITY OF THREE INFINITE SERIES IN CALCULATING AREA UNDER NORMAL PROBABILITY CURVE.



As an illustration of the use of this graph (Figure 1), we note that for five-decimal accuracy Series A must be used for all values of  $\frac{x}{\sigma}$  up to 2.50, and that Series S may be used for  $\frac{x}{\sigma} = 2.50$  and all larger values. Table 1 shows that the number of terms required in Series A for five-decimal accuracy increases from 3 at  $\frac{x}{\sigma} = .25$  to approximately 14 at  $\frac{x}{\sigma} = 2.25$ ; while, beginning at  $\frac{x}{\sigma} = 2.50$ , Series S requires but 4 terms, and this number diminishes to 1 term for  $\frac{x}{\sigma} = 5.00$ .

TABLE 1.

NUMBERS OF TERMS REQUIRED FOR VARYING DEGREES OF ACCURACY IN  
CALCULATION OF INCREASING PROPORTIONS OF AREA UNDER  
THE NORMAL PROBABILITY CURVE.

	3 decimals			4 decimals			5 decimals			7 decimals			9 decimals		
	A	B	S	A	B	S	A	B	S	A	B	S	A	B	S
.25	2	x	x	2	x	x	3	x	x	4	x	x	5	x	x
.50	3	x	x	3	x	x	4	x	x	5	x	x	6	x	x
.75	4	x	x	4	x	x	5	x	x	7	x	x	8	x	x
1.00	4	x	x	5	x	x	6	x	x	8	x	x	9	x	x
1.25	5	x	x	6	x	x	7	x	x	9	x	x	11	x	x
1.50	7	x	x	8	x	x	9	x	x	11	x	x	12	x	x
1.75	7	x	5	9	x	x	10	x	x	12	x	x	14*	x	x
2.00	9	x	3	10	x	4	12	x	x	14*	x	x	16*	x	x
2.25	11	x	3	12	x	3	14*	x	x	15*	x	x	18*	x	x
2.50	12	2	2	13*	x	2	15*	x	4	17*	x	x	20*	x	x
2.75	-	1	1	-	x	2	-	x	4	19*	x	x	23*	x	x
3.00	-	1	1	-	3	2	-	x	3	-	x	4	27*	x	x
3.50	-	1	1	-	1	1	-	3	2	-	x	4	-	x	7
4.00	-	1	1	-	1	1	-	1	1	-	5	3	-	x	6
5.00	-	1	1	-	1	1	-	1	1	-	1	1	-	3	2
	3 decimals			4 decimals			5 decimals			7 decimals			9 decimals		

\* Estimated by graphic extrapolation.

Explanatory: Read table as follows: The number of terms required in Series A to calculate to 4 decimal places of accuracy the portion of the area under the normal curve lying between the ordinates at  $\frac{x}{\sigma} = 0$  and  $\frac{x}{\sigma} = 2.00$  is 10; with Series S it is 4; the calculation is impossible to this degree of accuracy with Series B.

Notes: x indicates impossible calculation. - indicates impracticable calculation.

## CONCLUSIONS

All areas under the normal curve may be calculated by the use of Series A and Series S, the two being complementary.

Methods of developing Series A and Series B are outlined and it is indicated that Series S is derived from Series B.

The domain of practical utility for each series is shown in Figure 1. The numbers of terms required for various degrees of accuracy are shown in Table 1.

### SELECTED REFERENCES

1. Brunt, David, *The Combination of Observations*. Cambridge University Press, 1917.
2. Elderton, W. P., *Frequency Curves and Correlation*. Layton, London, 1927.
3. Kelley, T. L., *Statistical Method*. Macmillan, 1923.
4. Pearson, K., *Tables for Statisticians and Biometricians*. Part I, Second Edition (1924). Cambr. U. Press.
5. Rietz, H. L., Editor, *Handbook of Mathematical Statistics*. Houghton Mifflin Co., 1924.
6. Townsend, E. J., *Functions of Real Variables*. Holt, 1928.
7. Weld, L. D., *Theory of Errors and Least Squares*. Macmillan, 1916.
8. Wilson, E. B., *Advanced Calculus*. Ginn and Co., 1912.

## EDITORIAL: A. L. O'TOOLE

### ON A BEST VALUE OF $R$ IN SAMPLES OF $R$ FROM A FINITE POPULATION OF $N$ .

In recent years the problem of finding the moment coefficients for samples of  $n$  drawn from a finite population of  $N$  has been of interest to so many writers<sup>1</sup> that it seems worthwhile to make a few further observations<sup>2</sup> concerning these moment coefficients—particularly with respect to their dependence on  $n$ . In many instances the value of  $n$  to be used is at the discretion of the investigator and he would like to know if there is one value of  $n$  which is better than any other. An answer to that question will be given here.

---

<sup>1</sup> H. C. Carver, On the fundamentals of the theory of sampling, *Annals of Mathematical Statistics*, Vol. I, No. 1, pp. 101-121; Vol. I, No. 3, pp. 260-274.

C. C. Craig, An application of Thiele's semi-invariants to the sampling problem, *Metron*, Vol. VII, No. 4, 1928, pp. 3-74.

R. A. Fisher, Moments and product moments of sampling distributions, *Proc. London Math. Soc.*, Series 2, xxix, 1929, pp. 309-321; xxx, 1929, pp. 199-238.

L. Isserlis, On a formula for the product moment coefficients of any order of a normal frequency distribution in any number of variables. *Biometrika*, xii, 1918-19, pp. 134-139.

P. R. Rider, Moments of moments, *Proc. of the National Academy of Sciences*, Vol. 15, 1929, pp. 430-434.

H. E. Soper, Sampling moments of samples of  $n$  units each drawn from an unchanging sampled population, from the point of view of semi-invariants, *Journal of the Royal Statistical Soc.*, Vol. 93, 1930, pp. 104-114.

A. A. Tchouproff, On the mathematical expectation of the moments of frequency distributions, *Biometrika*, xii, 1918-19, pp. 140-169 and 184-210; xiii, 1920-21, pp. 283-295.

A. L. O'Toole, On symmetric functions and symmetric functions of symmetric functions, *Annals of Mathematical Statistics*, Vol. II, No. 2, May 1931, pp. 102-149. See Chapter III.

<sup>2</sup> These observations arose as a result of some very far-reaching suggestions on the theory of sampling made by Professor Carver, during recent conversations with the writer.

The differential operator method developed by this writer<sup>3</sup> for finding the moment coefficients not only was a very simple method but had the added advantage of leading directly to some theorems whose generality had not been established previously.

Using the notation of the previous paper let the finite parent population of  $N$  be composed of the  $N$  variates  $x_1, x_2, x_3, \dots, x_N$ . From this population draw all of the  ${}_N C_r$  different samples and let  $z_i = \sum_{k=1}^r x_k$ ,  $i = 1, 2, 3, \dots, {}_N C_r$ , where  $\sum_{k=1}^r x_k$  designates the sum of the  $r$  values of  $x$  which appear in the  $i^{\text{th}}$  sample. With this notation it has been shown in the paper cited that

$$(1) \quad S_{t; z} = t! \sum \frac{P_i^I \cdot P_j^J \cdot P_k^K \cdots S_{i;x}^I S_{j;x}^J S_{k;x}^K \cdots}{(i!)^I (j!)^J (k!)^K \cdots (I!)(J!)(K!) \cdots}$$

$$\text{where } S_{t; z} = \sum_{i=1}^{{}_N C_r} z_i^t, \quad t = 1, 2, 3, \dots$$

$$\text{and } S_{w; x} = \sum_{i=1}^N x_i^w, \quad w = 1, 2, 3, \dots$$

The summation in (1) is to be taken over terms such that  $Ii + Jj + Kk + \dots = t$  where  $I, J, K, \dots, i, j, k, \dots$  are positive integers, and where  $P_m$  is obtained from the  $m^{\text{th}}$  sampling polynomial  $P_m(p)$  by replacing the exponents of the polynomial by corresponding subscripts.

$$(2) \quad P_m(p) = \sum_{i=0}^{m-1} (-1)^i \binom{m-1}{i} p^{i+1}, \quad \text{OR}$$

$$(3) \quad P_m(p) = \left. \frac{d^m}{dx^m} \log (pe^x + 1 - p) \right|_{x=0}.$$

<sup>3</sup> Loc. cit.

In particular  $P_1 = p_1$ ,  $P_2 = p_1 - p_2$ ,  $P_3 = p_1 - 3p_2 + 2p_3$

$$P_4 = p_1 - 7p_2 + 12p_3 - 6p_4$$

$$P_5 = p_1 - 15p_2 + 50p_3 - 60p_4 + 24p_5$$

$$P_6 = p_1 - 31p_2 + 180p_3 - 390p_4 + 360p_5 - 120p_6$$

$$P_7 = p_1 - 63p_2 + 602p_3 - 2100p_4 + 3360p_5 - 2520p_6 + 720p_7$$

where  $p_K = {}^{N-K}C_{N-K}$ ,  $K \leq N$ .

It must be kept in mind that the multiplication of these operators is symbolic. For example, to find  $P_i^I P_j^J$  first multiply the polynomials  $P_i^I(p)$  and  $P_j^J(p)$  by ordinary multiplication and then the result when the exponents in this product are replaced by corresponding subscripts is  $P_i^I P_j^J$ .

Since in this paper it is desired to consider moments rather than power sums, replace  $s_{t;z}$  by  $({}^N C_N) \mu'_{t;z}$  and  $s_{w;x}$  by  $N \mu'_{w;x}$ . Then (1) becomes, after dividing by  ${}^N C_N$ ,

$$(4) \quad \mu'_{t;z} = \frac{t!}{{}^N C_N} \sum \frac{P_i^I P_j^J P_k^K \cdots N^{I+J+K}}{(i!)^I (j!)^J (k!)^K \cdots} \frac{\mu'_{i;z}^I \mu'_{j;z}^J \mu'_{k;z}^K \cdots}{I! \cdot J! \cdot K! \cdots}$$

Now  $p_k = {}^{N-k}C_{N-k}$ ,  $k \leq N$ , hence,

$$\frac{p_k}{{}^N C_N} = \frac{N(N-1)(N-2) \cdots (N-k+1)}{N(N-1)(N-2) \cdots (N-k+1)}$$

Substituting this value for each  $p_k/{}^N C_N$  in (4) the result is

### Equations 5.

$$\mu'_{1;z} = N \mu'_{1;x}$$

$$\mu'_{2;z} = \frac{N}{N-1} [N(N-1) \mu_{1;x}^{\prime 2} + (N-N) \mu'_{2;x}]$$

$$\mu'_{3;z} = \frac{N}{(N-1)(N-2)} [N^2(N-1)(N-2) \mu_{1;x}^{\prime 3} + 3N(N-1)(N-N) \mu'_{1;x} \mu'_{2;x} + (N-N)(N-2N) \mu_{3;x}']$$

$$\mu'_{4;z} = \frac{n}{(n-1)(n-2)(n-3)} \left[ \begin{aligned} & n^3 (n-1)(n-2)(n-3) \mu'_{1;x} \\ & + 6 n^2 (n-1)(n-2)(n-3) \mu'^2_{1;x} \mu'_{2;x} \\ & + 4 n (n-1)(n-2)(n-2n+1) \mu'_{1;x} \mu'_{3;x} \\ & + 3 n (n-1)(n-2)(n-1) \mu'^2_{2;x} \\ & + (n-2)(n^2 + 6 n n + 6 n^2 + n) \mu'_{4;x} \end{aligned} \right]$$

etc.

Now let  $n = ar$ . ThenEquations 6.

$$\mu'_{1;z} = r \mu'_{1;x}$$

$$\mu'_{2;z} = \frac{r^2}{ar-1} \left[ a(n-1) \mu'^2_{1;x} + (a-1) \mu'_{2;x} \right]$$

$$\mu'_{3;z} = \frac{r^3}{(ar-1)(ar-2)} \left[ \begin{aligned} & a^2(n-1)(n-2) \mu'^3_{1;x} + 3a(n-1)(a-1) \mu'_{1;x} \mu'_{2;x} \\ & + (a-1)(a-2) \mu'_{3;x} \end{aligned} \right]$$

$$\mu'_{4;z} = \frac{r^3}{(ar-1)(ar-2)(ar-3)} \left[ \begin{aligned} & a^3 r (n-1)(n-2)(n-3) \mu'^4_{1;x} \\ & + 6 a^2 r (n-1)(n-2)(a-1) \mu'^2_{1;x} \mu'_{2;x} \\ & + 4 a (n-1)(a-1) \{ r(a-2) + 1 \} \mu'_{1;x} \mu'_{3;x} \\ & + 3 a (n-1)(a-1) \{ r(a-1) - 1 \} \mu'^2_{2;x} \\ & + (a-1) \{ r(a^2 - 6a + 6) + a \} \mu'_{4;x} \end{aligned} \right]$$

etc.



A partial check at this point is to note that for  $a = 1$  only the first term of each of these moment coefficients remains.

Let  $a = 2$ . Then the above moment coefficients become

$$(7) \left\{ \begin{aligned} \mu'_{1;\bar{x}} &= n \mu'_{1;x} \\ \mu'_{2;\bar{x}} &= \frac{n^2}{2n-1} \left[ 2(n-1) \mu'^2_{1;x} + \mu'_{2;x} \right] \\ \mu'_{3;\bar{x}} &= \frac{n^3}{2n-1} \left[ 2(n-2) \mu'^3_{1;x} + 3 \mu'_{1;x} \mu'_{2;x} \right] \\ \mu'_{4;\bar{x}} &= \frac{n^3}{(2n-1)(2n-3)} \left[ 4n(n-2)(n-3) \mu'^4_{1;x} + 12n(n-2) \mu'^2_{1;x} \mu'_{2;x} \right. \\ &\quad \left. + 4 \mu'_{1;x} \mu'_{3;x} + 3(n-1) \mu'^2_{2;x} - \mu'^4_{1;x} \right] \\ \mu'_{5;\bar{x}} &= \frac{n^4}{(2n-1)(2n-3)} \left[ n(n-3)(n-4) \mu'^5_{1;x} + 20n(n-3) \mu'^3_{1;x} \mu'_{2;x} \right. \\ &\quad \left. - 20 \mu'^2_{1;x} \mu'_{3;x} + 15(n-1) \mu'_{1;x} \mu'^2_{2;x} \right. \\ &\quad \left. - 5 \mu'_{1;x} \mu'_{4;x} \right] \end{aligned} \right.$$

etc.

It is observed that when  $a = 2$ , i.e. when  $n = 2n$ , the moment coefficient  $\mu'_{3;\bar{x}}$  is independent of the moment coefficient  $\mu'_{5;x}$ . Also  $\mu'_{5;\bar{x}}$  is independent of  $\mu'_{7;x}$ . But one must not assume that all the odd moment coefficients of  $\bar{x}$  are independent of the corresponding odd moment coefficients of  $x$ . For  $\mu'_{7;\bar{x}}$  is not independent of  $\mu'_{7;x}$  as is seen by evaluating  $P_7$  which is the coefficient of  $\mu'_{7;x}$  in the expression for  $\mu'_{7;\bar{x}}$ .

So far the moments considered have been the moments of  $x$  with respect to the origin from which  $x$  is measured and the moments of  $\bar{x}$  with respect to the origin from which  $\bar{x}$  is measured. Consider now the moments of  $x$  about the mean value of  $x$  and the moments of  $\bar{x}$  about the mean value of  $\bar{x}$ . That is let

$$\begin{aligned} \bar{z}_i &= \bar{x}_i - M_{\bar{x}}, & i &= 1, 2, 3, \dots, n^{C_2}; \\ \bar{x}_i &= x_i - M_x, & i &= 1, 2, 3, \dots, N. \end{aligned}$$

Then

$$\begin{aligned}\bar{z}_i &= z_i - M_z = \sum_{j=1}^{n:i} x_j - n M_x \text{ since } M_z = n M_x \text{ by (5),} \\ &= \sum_{j=1}^{n:i} (x_j - M_x) = \sum_{j=1}^{n:i} \bar{x}_j.\end{aligned}$$

$$\begin{aligned}\text{e.g. } \bar{z}_1 &= z_1 - M_z = x_1 + x_2 + x_3 + \dots + x_n - n M_x \\ &= (x_1 - M_x) + (x_2 - M_x) + (x_3 - M_x) + \dots + (x_n - M_x) \\ &= \bar{x}_1 + \bar{x}_2 + \bar{x}_3 + \dots + \bar{x}_n \\ &= \sum_{j=1}^{n:1} \bar{x}_j.\end{aligned}$$

Hence it is clear that  $\bar{z}$  is the same function of  $\bar{x}$  as  $z$  is of  $x$ . In other words  $\mu_{z;\bar{z}}$  — (the moment of  $\bar{z}$  about the mean of  $\bar{z}$ ) — is the same function of  $\mu_{1;x}, \mu_{2;x}, \mu_{3;x}, \dots$  — (the moments of  $x$  about the mean of  $x$ ) — as  $\mu'_{z;z}$  was of  $\mu'_{1;x}, \mu'_{2;x}, \dots$ . There is one important simplification however due to the fact that  $\mu_{1;x} = 0$  and hence all terms which involve  $\mu_{1;x}$  vanish. With this in mind (6) becomes

$$(8) \left\{ \begin{aligned} \mu_{1;\bar{z}} &= 0, & \mu_{2;\bar{z}} &= \frac{n^2(a-1)}{a n - 1} \mu_{2;x}, \\ \mu_{3;\bar{z}} &= \frac{n^3(a-1)(a-2)}{(a n - 1)(a n - 2)} \mu_{3;x} \\ \mu_{4;\bar{z}} &= \frac{3 a n^3(n-1)(a-1)[n(a-1)-1]}{(a n - 1)(a n - 2)(a n - 3)} \mu_{2;x}^2 \\ &\quad + \frac{n^4(a^3 - 7a^2 + 12a - 6) + n^3 a(a-1)}{(a n - 1)(a n - 2)(a n - 3)} \mu_{4;x} \\ \mu_{5;\bar{z}} &= \frac{10 a(a-1)(a-2)(n-1)(a n - n - 1)}{(a n - 1)(a n - 2)(a n - 3)(a n - 4)} \mu_{2;x} \mu_{3;x} \\ &\quad + \frac{n^4(a-1)(a-2)(a^2 n - 12 a n + 12 n + 5 a)}{(a n - 1)(a n - 2)(a n - 3)(a n - 4)} \mu_{5;x} \\ &\quad \text{etc.} \end{aligned} \right.$$

Here again it is noticed that for  $a = 2$ , i.e. for  $N = 2n$ ,  $\mu_{3;\bar{z}}$  is independent of  $\mu_{3;x}$ . In other words the skewness of the distribution of  $\bar{z}$  is independent of the skewness of the parent

population of  $x$ . Similarly  $\mu_{5;\bar{x}}$  is independent<sup>4</sup> of  $\mu_{5;x}$  and also independent of  $\mu_{3;x}$ . But since  $P_7$  is not zero for  $a = 2$ ,  $\mu_{7;\bar{x}}$  is not independent of  $\mu_{7;x}$ .

Now consider the variance of  $\bar{x}$ ,

$$\begin{aligned}\mu_{2;\bar{x}} &= \frac{n^2(a-1)}{an-1} \mu_{2;x} \\ &= \frac{Nn - n^2}{N-1} \mu_{2;x} \quad (\text{since } N = an).\end{aligned}$$

Obviously it would be very desirable to have the variance (squared standard deviation) a minimum. Since the variance is a function of  $n$  differentiate  $\mu_{2;\bar{x}}$  with respect to  $n$ .

$$\frac{d}{dn} \mu_{2;\bar{x}} = \frac{N-2n}{N-1} \mu_{2;x}.$$

To make  $\mu_{2;\bar{x}}$  a minimum  $\frac{N-2n}{N-1} \mu_{2;x} = 0$  and hence  $N = 2n$  or, that is,  $a = 2$ .

$$\text{When } a = 2, \quad \mu_{2;\bar{x}} = \frac{n^2}{4(N-1)} \mu_{2;x}, \quad \sigma_{\bar{x}} = \frac{N}{2\sqrt{N-1}} \sigma_x.$$

In conclusion it may be said that there would seem to be good reason to suggest that, when possible, the investigator arrange to have twice as many variates in the control group or parent population as in each of the samples to be analyzed. Taking  $n = \frac{N}{2}$  will insure that the skewness of the samples will be independent of the skewness of the parent population and also that the fifth moment of the samples will be independent of the fifth moment of the sampled population. In addition, taking  $n = \frac{N}{2}$  will cause the variance (squared standard deviation) of the samples to be a minimum. Choosing  $n = \frac{N}{2}$  presumes, of course, that  $N$  is an even number. But in most instances it should be possible to arrange that  $N$  be even. For if an odd number of observations are given either another observation may be added or one of the given observations deleted to make  $N$  even.

<sup>4</sup>  $P_5$  vanishes with  $P_3$  because  $P_5 = P_3(1 - 12P_2)$ . But  $P_3$  is not a factor of  $P_7$ .

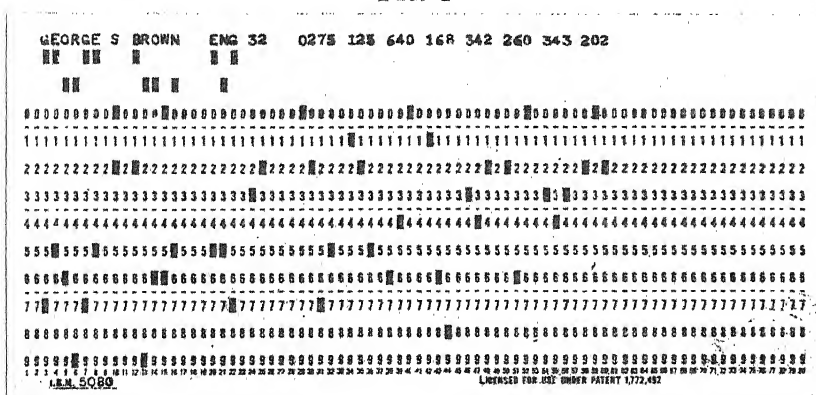
## EDITORIAL: H. C. CARVER

### PUNCHED CARD SYSTEMS AND STATISTICS

Because of the increasingly important part being played by mechanical devices in statistical methodology, it seems desirable to call attention in the *Annals* to some of the possibilities of punched-card systems.

The standard punched-card, illustrated below, is seven and one-half inches by three and one-quarter inches in size. To a certain extent the operation of a punched-card system is analogous to that of the Teletype machines used in wiring messages. In the latter case telegrams are written on a special typewriter which translates the message into a series of electrical impulses that in turn operate a distant typewriter which prints the message on a strip of paper; in the former, cards are automatically fed through a special typewriter that both prints the words or numbers on each card and also translates the information into properly punched holes in the card. The data on these cards may be totaled if desired by running the cards through a tabulating

Fig. 1



machine at a rate exceeding one per second; the total of the squares of the numbers appearing on consecutive cards may be obtained automatically; if the variates  $x$  and  $y$  be punched in respective columns on each card, the total of the  $xy$  products for all the cards is likewise made available; the cards may be arranged in order of magnitude according to card number, date, or variates at a rate exceeding six per second, and finally the data on the cards may be printed on a scroll—the cards passing through the listing or printing machine at about 80 per minute.

In order to provide an actual problem to serve as an illustration, I secured the anthropometric records of one thousand first year male students who entered the University of Michigan in the fall of 1928. The 1,000 cards, of which Figure 1 is a sample, were punched by an operator of average ability in slightly less than three hours. The data for the students were selected at random and the cards, as punched, were numbered consecutively from 1 to 1,000,—the card selected for Figure 1 being the 275th. The weight to the nearest pound of this individual was 125 pounds, and the height, width of shoulders, and the circumferences of chest, waist, hips and right thigh were, respectively, 64.0, 16.8, 34.2, 26.0, 34.3 and 20.2,—linear measurements being made to the nearest one-tenth inch.

These 1,000 cards may now be placed in a *tabulating and listing* machine which can total all seven of the data fields simultaneously. If desired, the machine will also print all of the information of each card on a scroll together with the totals. The first and last parts of this scroll are reproduced below photographically,—the names of the individuals being omitted purposely. Because of both the number of columns involved and the magnitudes of the totals, the listed totals unfortunately run together. The vertical lines, inserted with a pen, facilitate the reading of the totals. The cards are totaled and listed simultaneously at the rate of 80 per minute.

1	154	681	165	335	285	365	195
2	121	674	164	342	265	387	190
3	158	677	177	365	306	382	222
4	134	692	166	350	266	345	196
5	138	698	165	355	287	346	195
6	143	712	162	366	272	360	200
7	166	683	171	377	225	343	222
8	135	682	167	365	280	344	196
9	145	691	176	372	300	361	202
10	115	645	157	328	248	330	180

991	125	658	158	360	253	342	192
992	149	684	160	340	284	364	211
993	138	664	162	352	281	354	190
994	149	698	168	353	285	370	212
995	128	630	176	365	280	345	205
996	133	625	167	367	267	353	191
997	130	667	168	370	278	345	200
998	142	718	166	343	275	370	201
999	141	680	174	370	300	360	212
1000	135	686	174	345	280	365	206

1392886|678899|165337|353241|281510|3651632|01096

Fig. 2

An investigation of the correlation that may exist between height and weight will involve the numerical value of

$$\sum_{i=1}^{1000} x_i y_i$$

where  $x_i$  and  $y_i$  designate the height and weight, respectively, of the  $i^{\text{th}}$  individual. The plugboard of an *Automatic Multiplying Punch* may be wired in a few seconds so that

- the data of columns 34, 35 and 36 of Figure 1 will feed into the multiplier of the punch,
- the data of columns 38, 39 and 40 will feed into the multiplicand, and then
- the product,  $x_i y_i$ , for any card run through the machine will appear on the *product summary counter*. As the cards pass through the machine, the total of the products is accumulated on this counter.

If desired, each product may be punched automatically in the card, provided of course the card contains a sufficiently large number of otherwise vacant columns. The maximum number of digits in current models that may occur in either multiplier or multiplicand is eight. The number of digits in the multiplicand does not affect the speed of the multiplication; for three or less digits in the multiplier the cards feed through the machine at the rate of three seconds per card,—for eight digits in the mul-

tiplier the speed is twelve cards per minute. One may therefore place our cards in the machine, press a button, resume other duties, and some fifty minutes later the 1,000 cards will have yielded the total

$$\Sigma xy = 9\,477\,433.6 .$$

To obtain the sum of the squares of the variates in question it is necessary only to double-wire the machine,—one wire going to the multiplier and the other to the multiplicand. We obtain then

$$\Sigma x^2 = 4\,615\,312.12 \quad \Sigma y^2 = 19\,692\,452 .$$

By permitting the machine to punch each value of  $x$  in the card, we may treat  $x^2$  as the multiplicand and  $x$  as the multiplier and then obtain the sum of the cubes of the variates; or by double-wiring  $x^2$  obtain the sum of the fourth powers of the variates. If, while accumulating the cubes of the variates, we let the machine also punch each cube in the card, we may then obtain the sum of the powers of the variates up to and including that of the sixth order, etc. We are limited, of course, by the fact that the card contains eighty columns.

By running the punched cards through a sorting machine, we may obtain very readily the frequency distribution of the weights, and also the corresponding median, quartiles, etc. To accomplish this the cards must be run through a *sorting machine* three times, first sorting to column 35 of Figure 1, then to column 34 and finally to column 33. The cards pass through the sorting machine at the rate of 400 per minute, so that in approximately eight minutes—including time spent in handling the cards between sorts—these 1,000 cards will be perfectly arranged according to magnitude in weight. If the numbers with respect to which the sort is to be made contain  $n$  digits, the cards must be run through the sorter  $n$  times. We reproduce on the following page a photograph of the first part of a printed scroll obtained by running the cards through the listing machine after they had been sorted according to weight.

Figure 3.

Cand #	Weight	Height	Shoulder	Chest	Waist	Hip	R. High
232	89	597	140	295	253	305	171
146	100	607	157	300	242	312	170
691	101	600	150	301	245	308	165
358	102	676	154	327	262	322	174
555	102	662	153	303	242	307	165
941	102	640	154	313	262	313	168
209	103	663	154	310	285	313	172
801	103	649	150	302	247	317	168
14	104	638	147	313	253	300	178
513	105	635	148	310	250	328	187
720	105	637	155	330	258	319	178
563	106	648	152	302	245	320	169
672	106	638	162	323	254	320	181
153	107	623	143	317	250	322	172
75	108	630	160	335	245	332	183
235	108	669	157	320	247	324	172
322	108	624	165	305	260	325	188
505	108	637	152	330	246	330	183
31	109	620	161	334	265	340	192
393	109	625	152	314	233	336	187
30	110	631	162	345	265	327	180
160	110	667	153	320	247	322	175
185	110	627	152	323	274	320	172
631	110	630	153	335	255	330	180
802	110	623	149	317	257	328	181
15	111	650	158	312	265	344	194
151	111	637	157	322	258	325	172
273	111	651	156	332	270	326	180
447	111	655	160	314	252	324	176
20	112	696	155	300	230	335	174
426	112	647	154	333	262	328	170
507	112	691	161	323	262	322	180
716	112	667	148	330	254	322	185
831	112	651	147	373	262	333	182
308	113	661	155	318	250	330	180
383	113	665	154	338	240	333	178
449	113	651	151	328	258	311	163
541	113	675	152	316	257	340	186
591	113	650	166	341	265	333	180
826	113	654	147	331	260	333	183
898	113	677	156	320	253	335	187
933	113	656	153	315	240	326	179
947	113	637	146	315	262	337	188
967	113	696	151	320	258	330	169
148	114	654	153	320	247	322	183
177	114	633	160	342	255	321	172
257	114	653	162	318	270	320	185
368	114	679	145	318	243	322	173
462	114	653	153	322	261	321	171
464	114	627	150	310	262	332	180
545	114	612	157	334	260	318	181
741	114	662	161	338	262	338	186
951	114	681	153	311	242	333	174
987	114	667	152	342	250	330	188
10	115	645	157	328	248	330	180
139	115	665	146	332	265	339	188
159	115	636	160	328	297	318	174
336	115	663	153	321	252	323	167
743	115	639	158	322	278	328	182
949	115	686	154	330	242	330	182
970	115	653	152	313	250	337	172
988	115	692	155	330	252	352	172
52	116	640	157	316	260	337	193
186	116	606	159	330	258	337	186
226	116	665	166	341	262	335	185
268	116	636	165	350	265	350	192
347	116	630	163	321	265	325	195
510	116	662	162	337	275	330	190
523	116	695	158	325	243	340	182
926	116	641	154	338	257	340	182



A rough notion of the functional dependence that exists between weight and the other six variables recorded on the cards may be obtained by permitting the machine to total these ordered-with-respect-to-weight cards in consecutive groups of 100. That is, we obtain the averages for numerically equal groups selected according to the weight-deciles. The six regression lines may therefore be plotted, approximately, from the following results:

TABLE 1.  
ANTHROPOMETRIC AVERAGES BASED ON WEIGHT DECILES.

<i>Inter-decile Range</i>	<i>Weight</i>	<i>Height</i>	<i>Sk'der</i>	<i>Chest</i>	<i>Waist</i>	<i>Hips</i>	<i>Rt.Th.</i>
First	112.98	65.133	15.576	32.607	25.748	32.968	18.133
Second	122.41	66.659	15.980	33.663	26.777	33.927	18.926
Third	127.85	67.087	16.161	34.421	26.978	34.387	19.206
Fourth	131.98	67.381	16.334	34.816	27.622	34.858	19.554
Fifth	135.62	67.937	16.406	34.860	27.954	35.081	19.893
Sixth	139.54	68.189	16.651	35.608	28.065	35.511	20.112
Seventh	143.87	68.576	16.789	35.766	28.513	36.006	20.438
Eighth	149.43	68.895	16.807	36.116	28.780	36.420	20.712
Ninth	156.01	69.185	17.022	36.788	29.537	37.181	21.444
Tenth	173.19	69.854	17.611	38.596	31.536	38.826	22.678

If we had arranged the cards numerically with respect to height, instead of weight, we would have obtained the following results:

TABLE 2.  
ANTHROPOMETRIC AVERAGES BASED ON HEIGHT DECILES

<i>Inter-decile Range</i>	<i>Weight</i>	<i>Height</i>	<i>Sk'der</i>	<i>Chest.</i>	<i>Waist</i>	<i>Hips</i>	<i>Rt.Th.</i>
First	123.95	63.339	16.036	34.201	27.371	34.217	19.592
Second	130.97	65.295	16.261	34.856	27.984	34.944	19.929
Third	133.75	66.367	16.282	34.787	27.731	35.152	19.926
Fourth	136.28	67.021	16.570	35.429	28.329	35.302	20.084
Fifth	139.81	67.623	16.587	35.379	28.206	35.499	20.223
Sixth	140.60	68.189	16.532	35.510	28.184	35.560	20.008
Seventh	142.65	68.806	16.659	35.550	28.380	35.821	20.315
Eighth	143.44	69.498	16.638	35.328	28.065	35.858	20.205
Ninth	145.71	70.412	16.776	35.778	28.236	35.960	20.068
Tenth	155.72	72.346	16.996	36.423	29.024	36.852	20.746

# THE METHOD OF PATH COEFFICIENTS

By

SEWALL WRIGHT

Department of Zoology, The University of Chicago.

## Introduction

The method of path coefficients was suggested a number of years ago (Wright 1918, more fully 1920, 1921), as a flexible means of relating the correlation coefficients between variables in a multiple system to the functional relations among them. The method has been applied in quite a variety of cases. It seems desirable now to make a restatement of the theory and to review the types of application, especially as there has been a certain amount of misunderstanding both of purpose and of procedure.

## Basic Formulae

The object of investigation is a system of variable quantities, arranged in a typically branching sequential order representative of some chosen point of view toward the functional relations. Such a system is conveniently represented in a diagram such as Fig. 1. Those variables which are treated as 'dependent' are connected with those of which they are considered functions by arrows. The system of factors back of each variable may be made formally complete by the introduction of symbols representative of total residual determination (as  $V_0$  in Fig. 1). A residual correlation between variables is represented by a double-headed arrow. It will be assumed that all relations are linear.<sup>1</sup> Thus each variable is related to those from which uni-

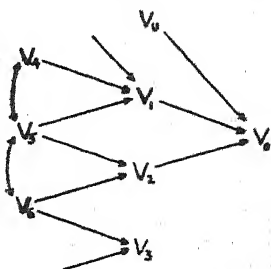


FIG. 1

<sup>1</sup> Relations which are far from linear with respect to the absolute values of the variables may be approximately linear with respect to variations, if the coefficients of variability are small. Thus if  $V_0 = f(V_1, V_2, V_3)$ ,

directional arrows are drawn to it by an equation of the following type, where  $(V_0 - \bar{V}_0)$ ,  $(V_1 - \bar{V}_1)$ , etc. represent deviations from the means and  $c_{01}$ ,  $c_{02}$  etc. are the coefficients.

$$(1) \quad (V_0 - \bar{V}_0) = c_{01}(V_1 - \bar{V}_1) + c_{02}(V_2 - \bar{V}_2) + \dots + c_{0n}(V_n - \bar{V}_n).$$

It is convenient to measure the deviation of each variable by its standard deviation. Let  $X_0 = \frac{V_0 - \bar{V}_0}{\sigma_0}$ ,  $X_i = \frac{V_i - \bar{V}_i}{\sigma_i}$  etc., and let  $P_{0i} = c_{0i} \frac{\sigma_i}{\sigma_0}$ ,

$$(2) \quad X_0 = P_{01} X_1 + P_{02} X_2 + \dots + P_{0n} X_n.$$

The coefficients in this form are of the type called path coefficients. Each obviously measures the fraction of the standard deviation of the dependent variable (with the appropriate sign) for which the designated factor is directly responsible, in the sense of the fraction which would be found if this factor varies to the same extent as in the observed data while all others (including residual factors  $X_u$ ) are constant. This definition (except for determination of sign) can be written as follows, putting the constant factors after a dot.

$$(3) \quad P_{0i} = \frac{\sigma_{0 \cdot 23 \dots n, u}}{\sigma_0} \cdot \frac{\sigma_i}{\sigma_{1 \cdot 23 \dots n, u}}.$$

It is sometimes convenient to represent the standard deviation due directly to a particular factor by a symbol. The form  $\sigma_{0(i)} = P_{0i} \sigma_0$  will be used. Obviously  $\sigma_{0(i)} = c_{0i} \sigma_i$  and, neglecting sign,

$$(4) \quad c_{0i} = \frac{\sigma_{0 \cdot 23 \dots n, u}}{\sigma_{1 \cdot 23 \dots n, u}}.$$

The theorem which makes the path coefficient useful in relating correlations to functional relation is a very simple one. The correlation between  $V_0$  and any other variable  $V_g$  in such a system as Fig. 2 can be written in the form

the relation of small deviations from the mean values are approximately the first order terms of an expansion by Taylor's Theorem. The error may be represented by a residual term  $R$ .

$$\delta V_0 = \frac{\partial V_0}{\partial V_1} \delta V_1 + \frac{\partial V_0}{\partial V_2} \delta V_2 + \dots + R.$$

$$\begin{aligned}
 (5) \quad r_{oq} &= \frac{1}{N} \sum X_o \cdot X_q = \frac{1}{N} \sum x_q (P_{o1} x_1 + P_{o2} x_2 + \dots + P_{on} x_n) \\
 &= P_{o1} r_{q1} + P_{o2} r_{q2} + \dots + P_{on} r_{qn} \\
 &= \sum P_{oi} \cdot r_{qi} .
 \end{aligned}$$

The correlation is thus analyzed into contributions from all of the paths in the diagram (Fig. 2) passing through each factor of one of the variables.

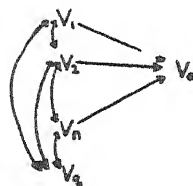


FIG. 2

But the correlation terms symbolized by  $r_{qi}$  may be capable of analysis by application of this same formula. By repeated analysis of this sort, as far as the diagram (such as Fig. 1) permits, we are led to the following principle: Any correlation between variables in a network of sequential relations can be analyzed into contributions from all of the paths (direct or through common factors) by which the two variables are connected, such that the value of each contribution is the product of the coefficients pertaining to the elementary paths. If residual correlations are present (represented by bidirectional arrows) one (but never more than one) of the coefficients thus multiplied together to give the contribution of a connecting path, may be a correlation coefficient. The others are all path coefficients.

In tracing connecting paths it is obvious that one may trace back along the arrows and then forward as well as directly from one variable to the other (perhaps through intervening variables) but never forward and then back. That two factors affect the same dependent variable does not contribute to the correlation between them. Similarly two variables which are correlated with a third are not necessarily correlated with each other. As illustrations of these principles consider the correlations between some of the variables in Fig. 1.

$$\begin{aligned}
 r_{36} &= P_{36} & r_{46} &= 0 & r_{13} &= P_{15} \cdot r_{56} \cdot P_{36} \\
 r_{23} &= P_{25} \cdot r_{56} \cdot P_{36} + P_{26} \cdot P_{36} & r_{12} &= P_{14} \cdot r_{45} \cdot P_{25} + P_{15} \cdot P_{25} + P_{15} \cdot r_{56} \cdot P_{26}
 \end{aligned}$$

It is sometimes convenient to use an extension of the symbolism in dealing with compound paths.

$$r_{o_2} = P_{o_2} + P_{o_1\dot{5}2} + P_{o_1\overline{4}52} + P_{o_1\overline{5}62}$$

$$r_{o_5} = P_{o_15} + P_{o_25} + P_{o_1\overline{4}5} + P_{o_2\overline{6}5}.$$

In this symbolism all of the variables along a contributing path are listed in proper order. If the path passes through a represented common factor, the latter is indicated by a dot. If it involves an unanalyzed correlation the two ultimate correlated variables may be indicated by a line as above. The evaluation of such compound path coefficients is obvious.

$$P_{o_15} = P_{o_1} P_{15}, \quad P_{o_1\dot{5}2} = P_{o_1} P_{15} P_{25}$$

$$P_{o_1\overline{4}52} = P_{o_1} P_{14} r_{45} P_{25}, \quad P_{o_1\overline{4}5} = P_{o_1} P_{14} r_{45}, \quad \text{etc.}$$

It is to be noted that the symbolism does not apply to the indicated variables in an absolute sense but is always to be understood as relative to a particular arrangement of the variables, i.e. to a particular point of view with respect to the functional relations.

A special case of equation (5) arises if one correlates a variable with itself, taking into account *all* factors (known and unknown)

$$(6) \quad r_{oo} = \sum P_{oi} r_{oi} = 1.$$

This may be put in a form which is usually more convenient by further analysis of  $r_{oi} = P_{oi} + \sum P_{oj} r_{ij}$

$$(7) \quad \sum P_{oi}^2 + 2 \sum P_{oi} P_{oj} r_{ij} = 1.$$

#### Degree of Determination

From the formula  $P_{oi}^2 = \frac{\sigma_{oi}^2}{\sigma_o^2}$  it is obvious that a squared path coefficient measures the portion of the variance of the dependent variable for which the independent variable is directly responsible, under the point of view adopted. The squared path coefficient may accordingly be called a coefficient of determination. Such coefficients were used before the term path coefficient was applied to the square root. (Wright 1918.)

The sum of the squared path coefficients is unity only in the case in which there are no correlations among the factors. It is necessary, therefore, to recognize additional terms measuring the changes in variance (positive or negative) due to correlated occurrence of the contributions of such factors ( $2 P_{oi} P_{oj} r_{ij}$ , etc. in equation 7). It is tempting to apportion determination among the factors by using the terms  $P_{oi} r_{oi}$  of equation (6) as measures of determination, and this has been done by some authors, e.g. Kirchevsky (1927) who independently reached a somewhat similar viewpoint on the interpretation of systems of correlated variables in other respects. No transparent meaning can be attached to such expressions (which may be negative). The term does not measure direct determination since it involves indirect connections between the variables. Neither does it measure total determination, direct and indirect. This is given by the squared correlation coefficient.

### *The Correlation between Linear Functions*

The most direct application of the method is in the estimation of the correlation between two variables which are functions (in part at least) of the same variables. Let  $V_s$  and  $V_T$  be two variables whose correlation is desired.

$$\begin{aligned}
 V_s &= c_s + c_{s1} V_1 + c_{s2} V_2 + \dots + c_{si} V_i \\
 V_T &= c_T + c_{T1} V_1 + c_{T2} V_2 + \dots + c_{Ti} V_i \\
 (8) \quad \sigma_s^2 &= \sum c_{si}^2 \sigma_i^2 + 2 \sum c_{si} c_{sj} \sigma_i \sigma_j r_{ij} \\
 \sigma_T^2 &= \sum c_{Ti}^2 \sigma_i^2 + 2 \sum c_{Ti} c_{Tj} \sigma_i \sigma_j r_{ij} \\
 P_{si} &= c_{si} \frac{\sigma_i}{\sigma_s}, \quad P_{Ti} = c_{Ti} \frac{\sigma_i}{\sigma_T} \\
 (9) \quad r_{sT} &= \sum P_{si} P_{Ti} + \sum P_{si} r_{ij} P_{Tj}
 \end{aligned}$$

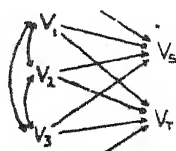


FIG. 3

As an example, suppose that we wish to find the correlation between the compound variables  $V_s$  and  $V_T$  where  $V_s = V_1 + V_2 + V_3$  and  $V_T = V_1 + V_2 + V_4$  knowing that  $V_1, V_2, V_3$  and  $V_4$  are all of equal variability ( $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4$ )

and are independent ( $r_{12} = r_{13} = r_{14} = r_{23} = r_{24} = r_{34} = 0$ )

$$\sigma_s^2 = \sigma_T^2 = 3 \sigma_i^2$$

$$P_{sT} = P_{s2} = P_{s3} = P_{T1} = P_{T2} = P_{T4} = \sqrt{1/3}$$

$$r_{sT} = P_{s1} P_{T1} + P_{s2} P_{T2} = 2/3.$$

Again suppose that we wish to estimate the true correlation between two variables from that between measurements known to be subject to considerable random error. Assume that the correlation between two measurements of the same variate has been found in each case. It is instructive to work this out from two different points of view.

Let  $\bar{A}$  be the mean of  $m$  measurements ( $A_1, A_2, \dots, A_m$ ) of  $A$ . Let  $\bar{B}$  be the mean of  $n$  measurements ( $B_1, B_2, \dots, B_n$ ) of  $B$ .

The known correlations are those between measures of  $A$  ( $r_{AA}$ ), between measures of  $B$  ( $r_{BB}$ ) and between measures of  $A$  and  $B$  ( $r_{AB}$ ). Expressing the complete determination of  $\bar{A}$  and  $\bar{B}$  by their components, using

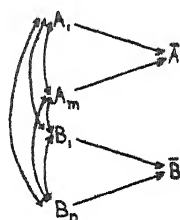


FIG. 4

$$\begin{aligned} \sum P_{\bar{A}A} r_{\bar{A}A} &= m P_{\bar{A}A}^2 [1 + (m-1) r_{AA}] = 1 \\ \sum P_{\bar{B}B} r_{\bar{B}B} &= n P_{\bar{B}B}^2 [1 + (n-1) r_{BB}] = 1 \end{aligned}$$

$$\text{From these} \quad P_{\bar{A}A} = \sqrt{\frac{1}{m[1 + (m-1)r_{AA}]}} \quad , \quad P_{\bar{B}B} = \sqrt{\frac{1}{n[1 + (n-1)r_{BB}]}}$$

The correlation between  $\bar{A}$  and  $\bar{B}$  can be written

$$\begin{aligned} r_{\bar{A}\bar{B}} &= m P_{\bar{A}A} r_{\bar{A}B} = m \cdot n \cdot P_{\bar{A}A} \cdot P_{\bar{B}B} \cdot r_{AB} \\ (10) \quad &= r_{AB} \sqrt{\frac{m \cdot n}{[1 + (m-1)r_{AA}][1 + (n-1)r_{BB}]}} \end{aligned}$$

For indefinitely large values of  $m$  and  $n$  the averages may be considered as true scores,  $A_T$  and  $B_T$ .

$$(11) \quad r_{A_TB_T} = \frac{r_{AB}}{\sqrt{r_{AA} \cdot r_{BB}}}$$

This result can be reached much more directly by the simpler set up (Fig. 5) in which the observed measurements are represented as functions of the true scores  $A_T, B_T$  and of random

errors. Note that the directions of the arrows are the reverse of those in Fig. 4.

$$r_{AA} = P_{AA_T}^2 \quad \text{giving} \quad P_{AA_T} = \sqrt{r_{AA}}$$

$$r_{BB} = P_{BB_T}^2 \quad \text{giving} \quad P_{BB_T} = \sqrt{r_{BB}}$$

$$r_{AB} = P_{AA_T} P_{A_T B_T} P_{BB_T} \quad \text{again giving} \quad r_{A_T B_T} = \frac{r_{AB}}{\sqrt{r_{AA} \cdot r_{BB}}}$$

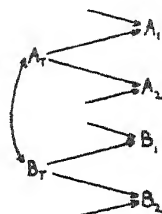


FIG. 5

This formula is, of course, Spearman's correction for attenuation. The purpose here is to bring out the simple way in which such formulae can be obtained by the method of path coefficients. The following is a more complex case in which a simple method is more essential.

### *The Statistical Effects of Inbreeding<sup>2</sup>*

Assume for simplicity that the effects of different genetic factors combine additively (no dominance or epistasis). In Fig. 6,  $P_1$  and  $P_2$  represent the genetic constitution of two parents and  $O$  of their offspring. The constitution of the latter (under the

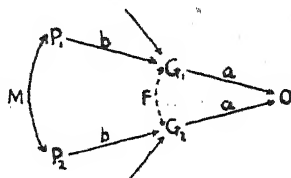


FIG. 6

above assumption and ignoring the possibility of sex linkage) is equally and completely determined by the constitutions of the two germ cells ( $G_1, G_2$ ) which united to produce it. It will be convenient to represent the path coefficients and correlations by single letters:

$$a = P_{OG_1} = P_{OG_2}, \quad b = P_{G_1 P} = P_{G_2 P},$$

$$F = r_{G_1 G_2}, \quad M = r_{P_1 P_2}.$$

The determination of  $O$  by  $G_1$  and  $G_2$  can be expressed in

<sup>2</sup> The purpose in presenting this and later examples is to illustrate something of the range of applicability of the method, rather than to give a detailed analysis of each case. For the latter the reader must be referred to the references cited at the end.



the equation  $2a^2(1+F) = 1$  (by equation 7) giving

$$(12) \quad a = \sqrt{\frac{1}{2(1+F)}}.$$

Two complementary germ cells ( $G_A, G_B$ ) Fig. 7, such as could arise from the same reduction division have the same relation to the genetic constitution of the parent (which they

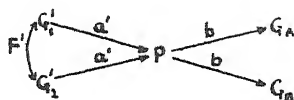


FIG. 7

completely determine in a mathematical sense) as the two germ cells which united to produce the parent, assuming no selection and that different series of allelomorphs are combined at random, (an assumption compatible with linkage among the genes and with inbreeding, but not with assortative mating). Using primes for the path coefficients and correlations of the preceding generations.

$$(13) \quad b = r_{PG'} = a'(1+F') = \sqrt{\frac{1+F'}{2}}.$$

Since  $a' = \sqrt{\frac{1}{2(1+F)}}$  (by equation 12, applied to the preceding generation)  $b a' = 1/2$  irrespective of correlation between the parents under the assumed conditions.

$$(14) \quad F = b^2 M, \quad M = \frac{2F}{1+F'}.$$

The correlation between uniting gametes is directly related to the percentage of heterozygosis. Below is the correlation table between uniting gametes in a population in which genes  $A$  and  $a$  are present in the frequencies of  $q$  and  $1-q$  respectively and the proportion of heterozygotes is  $p$ . By the usual formula for correlation

$$(15) \quad F = \frac{(q - 1/2) - q^2}{q(1-q)} = 1 - \frac{p}{2q(1-q)}$$

$$p = 2q(1-q)(1-F).$$

All of the path coefficients and correlations have now been expressed in terms of  $F$ 's. Various applications can be made. As a simple case consider the

	$a$	$A$	Total
$A$	$\frac{p}{2}$	$q - \frac{p}{2}$	$q$
$a$	$1 - q - \frac{p}{2}$	$\frac{p}{2}$	$1 - q$
Total	$1 - q$	$q$	1

effects of continued brother-sister mating: Analyzing the correlation ( $M$ ) Fig. 8, between the parents by tracing the connecting paths:

$$(16) \quad M = 2 a'^2 b'^2 (1 + M').$$

Expressing all coefficients in terms of  $F_s'$  and reducing

$$(17) \quad F = 1/4 (1 + 2F' + F'')$$

$$(18) \quad P = \frac{P'}{2} + \frac{P''}{4}.$$

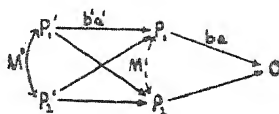


FIG. 8

Thus the percentage of heterozygosis (to which the effects of inbreeding are directly related) is very simply related to the percentages in the two preceding generations. If there were initially 50% i.e.  $\frac{1}{2}$  heterozygosis, that of later generations would be given by the terms of a series of fractions in which each numerator is the sum of the two preceding numerators (Fibonacci series) if the denominator is doubled in each generation. This rule was derived empirically by Jennings (1916) on working out in detail the consequences of every possible mating, generation after generation. The analysis by path coefficients (Wright 1921b) not only demonstrates the generality of the empirical rule but can be applied as easily to more complicated cases in which the analysis by types of mating would be practically impossible. Consider, for example, the more general case of a population restricted to  $N_m$  mature males and  $N_f$  mature females (Wright 1931b). Under random mating, the chance of a mating of full brother and sister is  $\frac{1}{N_m \cdot N_f}$ , of half brother and sister  $\frac{N_m + N_f - 2}{N_m \cdot N_f}$  and of less closely related individuals  $\frac{(N_m - 1)(N_f - 1)}{N_m \cdot N_f}$ .

The correlation between mating individuals is thus

$$(19) M = d^2 b^2 \left[ \frac{2+2M'}{N_m \cdot N_f} + \left( \frac{N_m + N_f - 2}{N_m \cdot N_f} \right) (1+3M') + \frac{(N_m - 1)(N_f - 1)}{N_m \cdot N_f} \cdot 4M' \right]$$

which yields on reduction

$$(20) \quad F = F' + \left( \frac{N_m + N_f}{8 N_m N_f} \right) (1 - 2F' + F'')$$

$$(21) \quad p = p' \left( \frac{N_m + N_f}{8 N \cdot N} \right) (2 p' - p'').$$

Equating  $\frac{P}{P'}$  to  $\frac{P'}{P''}$  gives  $(\frac{1}{8N_m} + \frac{1}{8N_f})$  as the approximate rate of reduction of heterozygosis per generation. The special case in which the population is equally divided between males and females ( $N_m = N_f = \frac{N}{2}$ ) gives  $\frac{1}{2N}$  as the rate of reduction, a figure recently verified by R. A. Fisher by a very different mode of analysis.

The method has also been applied in the much more complicated case of assortative mating based on somatic resemblance (Wright 1921b).

In the case of the irregular inbreeding encountered in live stock pedigrees (Wright 1922, 1923a), the basic formula of path coefficients leads immediately to the formula

$$(22) \quad F = \sum \left[ \left(\frac{1}{2}\right)^{N_s + N_d + 1} \cdot (1 + F_A) \right],$$

where  $N_s$  and  $N_d$  are the number of generations from sire and dam respectively to the common ancestor ( $A$ ) at the head of each connecting path. By appropriate sampling methods (Wright & McPhee, 1925) this formula can be used in the study of whole breeds. Closely allied is the formula for the genetic correlation between any two individuals ( $X, Y$ ). Letting  $N$  and  $N'$  be the generations from  $X$  and  $Y$  respectively to the common ancestor of any connecting path

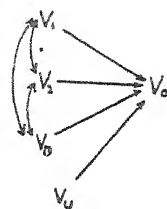
$$(23) \quad R_{XY} = \frac{\sum \left[ \left(\frac{1}{2}\right)^{N+N'} \cdot (1 + F_A) \right]}{\sqrt{(1 + F_X)(1 + F_Y)}}.$$

These formulae have been extensively applied in breed analysis (Wright 1923b-c, McPhee & Wright 1925, 1926, Smith 1926, Calder 1927, Lush 1932).

### *Multiple Regression*

The preceding applications have consisted in the main in the deduction of correlation coefficients from knowledge of the functional relations. The method can be applied as well to the inverse problem, that of finding the best linear expression for one variable in terms of a number of others, from knowledge of the cor-

relation coefficients. No assumptions are made with respect to causal relations. Analysis of the correlations between  $V_o$  and the other variables (Fig. 9), by the basic formula, gives the following set of equations.



$$\begin{aligned}
 r_{o1} &= P_{o1} + P_{o2} r_{12} + \dots + P_{on} r_{1n} \\
 r_{o2} &= P_{o1} r_{12} + P_{o2} + \dots + P_{on} r_{2n} \\
 &\dots \dots \dots
 \end{aligned}
 \tag{24}$$

FIG. 9

$$r_{on} = P_{o1} r_{1n} + P_{o2} r_{2n} + \dots + P_{on}$$

Obviously these are merely the normal equations of the method of least squares in a slightly disguised form, as might be expected from the derivation of the basic formula. The solution for the path coefficients, expressed in terms of determinants, merely need to be multiplied by the proper ratio of standard deviations to give Pearson's formulae for the partial regression coefficients. The method of path coefficients here merely furnishes a convenient mnemonic rule for writing the normal equations.

The correlation between the actual values of  $V_o$  and the estimates ( $V_o'$ ) (Fig. 10) from each set of values of the other variables, (given by the regression equation) is Pearson's coefficient of multiple correlation. Let  $V_u$  stand for the array of residual factors of  $V_o$  in a form independent of the known factors. We may write an equation of complete determination (Fig. 9)

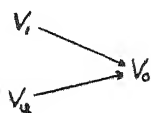


FIG. 10

$$\begin{aligned}
 \sum_i P_{oi} r_{oi} + P_{ou} r_{ou} &= 1 \\
 \sum_i P_{oi} r_{oi} &= 1 - r_{ou}^2 \quad \text{since } P_{ou} = r_{ou}
 \end{aligned}$$

But  $r_{oo'}^2 = 1 - r_{ou}^2$  (Fig. 10) Therefore

$$(25) \quad r_{oo'} = r_{o(12\dots n)} = \sqrt{\sum_i P_{oi} r_{oi}}$$

It is unnecessary to give illustrations of the use of the method in obtaining ordinary estimation or prediction equations.

A somewhat different type of application has been made in estimating the transmitting capacity of dairy sires (Wright 1932a). In this case the necessary correlations were deduced from Mendelian theory checked by observed correlations between the sire's female relatives and his daughters. These correlations were then used to calculate the multiple regression of sire on daughters and their dams.

### Partial Correlation

It is sometimes of interest to find the values which statistics would take, on the average, in data selected for constancy of one or more variables.

$$(26) \quad \sigma_{o,1,2,\dots,m}^2 = \sigma_{o(u)}^2 = \sigma_o^2 \cdot P_{ou}^2 = \sigma_o^2 (1 - r_{o(1,2,\dots,m)}^2),$$

a well known formula. Inspection of equations (3) and (4) and of the definition of  $\sigma_{o(1)}$  gives the following for the standard deviation of  $V_o$  due directly to  $V_1$ , under constancy of  $V_2$  etc., and for the related path coefficient and concrete partial regression coefficient under the same conditions.

$$(27) \quad \sigma_{o(1),2,\dots,m} = \sigma_{o(1)} \frac{\sigma_{1,2,\dots,m}}{\sigma_1} = \sigma_{o(1)} \sqrt{1 - r_{1(2,\dots,m)}^2}$$

$$(28) \quad P_{o1,2,\dots,m} = \frac{\sigma_{o(1),2,\dots,m}}{\sigma_{o,2,\dots,m}} = P_{o1} \sqrt{\frac{1 - r_{1(2,\dots,m)}^2}{1 - r_{o(2,\dots,m)}^2}}$$

$$(29) \quad r_{o1,2,\dots,m} = \frac{\sigma_{o(1),2,\dots,m}}{\sigma_{1,2,\dots,m}} = r_{o1}.$$

As might be expected, the concrete coefficients of the multiple regression equation ( $r_{o1}$ , etc.) remain the same (on the average) in samples selected for constancy of one or more of the factors, while the abstract path coefficients are altered in value in such samples.

The formula for partial correlation can be derived from the formula  $\sigma_{o,1}^2 = \sigma_o^2 (1 - r_{o,1}^2)$

as applied to the data in which particular variables ( $V_2 \dots V_m$ ) are constant.

$$(30) \quad \sigma_{0.12 \dots m}^2 = \sigma_{0.2 \dots m}^2 (1 - r_{01.2 \dots m}^2)$$

$$r_{01.2 \dots m}^2 = 1 - \frac{\sigma_{0.12 \dots m}^2}{\sigma_{0.2 \dots m}^2} = 1 - \frac{1 - r_{0(12 \dots m)}^2}{1 - r_{0(2 \dots m)}^2}.$$

This derivation leaves the sign uncertain but this is easily determined from a different approach. In Fig. 11,  $V_u$  includes all factors of  $V_o$  other than the designated independent variable  $V_i$  and the variables  $V_2 \dots V_m$  which are to be held constant.  $V_v$  represents the residual factor for  $V_i$ , in relation to the factors  $V_2 \dots V_m$ . The value of  $P_{o1}'$  in this system is not of course the same as  $P_{o1}$  in the preceding discussion in which other variables than  $V_2 \dots V_m$  (those to be made constant) were treated as factors of  $V_o$ .

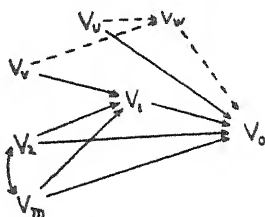


FIG. 11

Since

$$c_{01} = r_{01} \frac{\sigma_o}{\sigma_i}, \quad c_{01.2 \dots m} = r_{01.2 \dots m} \frac{\sigma_{0.2 \dots m}}{\sigma_{i.2 \dots m}}.$$

But  $c_{01.2 \dots m} = c_{01} = P_{o1} \frac{\sigma_o}{\sigma_i}$ . Therefore

$$(31) \quad r_{01.2 \dots m} = P_{o1} \cdot \frac{\sigma_o}{\sigma_i} \cdot \frac{\sigma_{i.2 \dots m}}{\sigma_{0.2 \dots m}} = P_{o1} \sqrt{\frac{1 - r_{i(2 \dots m)}^2}{1 - r_{0(2 \dots m)}^2}}.$$

Thus  $r_{01.2 \dots m}$  has the same sign as  $P_{o1}$ .

This is simply formula 28 except that it is in a set up in which all factors of  $V_o$  except  $V_i$  are held constant, in which case  $P_{01.2 \dots m}$  becomes  $r_{01.2 \dots m}$ .

$$\text{Since} \quad 1 - r_{0(12 \dots m)}^2 = P_{ou}^2$$

$$1 - r_{i(2 \dots m)}^2 = P_{iv}^2$$

$$1 - r_{0(2 \dots m)}^2 = P_{ou}^2 + P_{oi}^2 P_{iv}^2 \text{ OR } P_{ow}^2,$$

letting  $V_w$  represent the combination of  $V_u$  and  $V_v$  the above formulae for partial correlation can be written in a number of very compact forms.

$$(32) \quad r_{01.2 \dots m} = \sqrt{1 - \frac{P_{ou}^2}{P_{ow}^2}}$$

$$(33) \quad = \frac{P_{01} \cdot P_{1v}}{\sqrt{P_{ou}^2 + P_{01}^2 \cdot P_{1v}^2}}$$

$$(34) \quad = \frac{P_{01} \cdot P_{1v}}{P_{ow}}.$$

The first of these is identical with 30.

### Symbolism

The most widely current symbol for a partial regression coefficient is Yule's expression  $\ell_{01.2 \dots m}$ . Kelley (1923) uses a similar expression  $\beta_{01.2 \dots m}$  for the coefficients in abstract form. These have an advantage over the symbols used here ( $r_{01}$ ,  $P_{01}$  respectively) in that they define certain absolute functions of the variables, while the latter symbols have meaning only in relation to a particular arrangement. This relativity of meaning can not, however, cause confusion as long as one is dealing with only a single system. If the problem is of a more complex sort than the calculation of a prediction formula, the  $\beta$  symbolism becomes too cumbersome for convenience. The current symbolism has the further disadvantage of a certain lack of logical consistency. In the expression  $\sigma_{0.12}$ ,  $r_{01.23}$  and  $\ell_{01.23}$  the subscripts to the right of the dot are understood to represent factors held constant. In the expressions  $r_{0.12}$  and  $\beta_{01.23}$  this is not the case. If we wish to represent the multiple correlation of  $X_0$  with  $X_1$  and  $X_2$ , independent of  $X_3$ , or the beta (path coefficient) for the influence of  $X_1$  on  $X_0$  in data involving also  $X_2$  and  $X_3$  but in which  $X_3$  is held constant, it would apparently be necessary under the usual symbolism to write such ambiguous expressions  $r_{0.123.3}$  and  $\beta_{01.23.3}$  respectively. Pearson's method

of writing constant factors as subscripts to the left of the main symbol avoids these difficulties and is the one which I have followed in earlier papers. The dot symbolism has, however, become so firmly established in the cases of the standard deviation and correlation coefficient that it is probably best to recognize it as the general device for indicating constant factors and to replace it in those symbols in which it is used for a different purpose.

There is no difficulty in the case of multiple correlation. The expression  $r_{o(12)}$  may be used for the correlation of  $X_o$  with  $X_1$  and  $X_2$  jointly and the expression  $r_{o(12).3}$  is an unambiguous symbol for the multiple correlation independent of  $X_3$ .

As noted above, it is not desirable in the usual application of path coefficients to encumber the symbols with a list of the factors of which each dependent variable is treated as a function. This can be left to a diagram. Where a complete formal symbolism is desirable, the list of factors might follow a semicolon instead of a dot. Thus  $P_{o1.23.3}$  would unambiguously represent the path coefficient relating  $X_o$  to  $X_1$  in a system in which  $X_o$  is treated as a function of  $X_1$ ,  $X_2$  and  $X_3$  but in which  $X_3$  is to be held constant. There is, however, little need for such complicated expressions.

#### *Quantitative Evaluation of Causal Relations*

While the method of path coefficients is directly applicable to such problems as the estimation of correlation coefficients from knowledge of the mathematical relations between variables, or the converse (multiple regression) it was developed primarily as a means of combining the quantitative information given by a system of correlation coefficients with such information as may be at hand with regard to the causal relations, and thus of making quantitative an interpretation which would otherwise be merely qualitative.

How far such causal analysis has meaning is a question on which there is difference of opinion. Some authors (Pearson, Niles) have contended that the designation of the relation be-



tween two variables as one of cause and effect involves a false conception; that we can merely observe more or less perfect correlation. This view seems to imply that direction in time is of no significance, and indeed G. N. Lewis has recently argued for the complete symmetry of the physicist's time. The common sense view that direction in time is a basic perception is not without support, however.

Under the theory of relativity, the elementary physical reality seems to be the point event located at a particular position in the space and time of a particular viewpoint. The objective world is to be thought of as a complex network of point events. Although two such events sufficiently remote from each other in space, relative to their separation in time, may have their order of succession in time reversed in the systems of two different observers, order in time is invariant along any strand of this network involving continuity of physical action. Thus the succession of collisions suffered by a particular body or by a beam of light is the same to all observers. Such successions of events as involved in the movement of a shadow over a surface may indeed be reversed by change of viewpoint, if the shadow happens to be moving more rapidly than the velocity of light, but the continuity of physical action here is not along the path of the shadow but traces separately to each point in this path from the points of interception of the light. There is frequently difficulty in complex cases in distinguishing lines of direct causation from correlations due to common causation but in principle the distinction is clear enough. Experimental intervention is possible only in the true lines of causation.

In the world of large scale events, certain patterns tend to recur. Certain recurrent successions of events come to be recognized, experimentally or otherwise, as lines of causation in the above sense. Different lines of this character may come together in a certain type of event or may diverge from one. In many cases a fairly adequate representation of the course of nature can

be obtained by viewing it as a coarse network in which the "events" of interest are the deviations in the values of certain measurable quantities. A qualitative scheme depends on observation of sequences and experimental intervention. It is of interest to make such a scheme at least roughly quantitative in the sense of evaluating the relative importance of action along different paths. This was the primary purpose of the method of path coefficients.

### *Birth Weight of Guinea Pigs*

The simplest application of this sort has been in connection with the factors which determine the weight of guinea pigs at birth (Wright 1921a). Minot (1891) noted that the average birth weight is smaller, the greater the size of the litter. He reasoned that this might be due either to a competition between the developing foetuses, or merely to an effect of a large litter in stimulating somewhat premature birth. In confirmation of the latter hypothesis he found that the gestation period was several days shorter in large litters than in small ones and that there was in fact a direct relation between length of gestation period and birth weight. After some discussion, he concluded that the data afforded no evidence of growth competition and thus he decided in favor of the second hypothesis. I was able to confirm Minot's observations, obtaining the following data in a large stock of guinea pigs. The mean birth weight (in grams) of the animals in the litter is the birth weight used. The interval between litters, where less than 75 days is approximately the gestation period. Standard errors are given.

	Mean	S D	
B (Birth weight)	$82.24 \pm 0.51$	$18.60 \pm 0.36$	$r_{BI} = +.533 \pm .020$
I (Interval)	$68.93 \pm 0.05$	$1.91 \pm 0.04$	$r_{BL} = -.658 \pm .010$
L (Size of litter)	$2.91 \pm 0.04$	$1.29 \pm 0.03$	$r_{IL} = -.457 \pm .022$

The correlation between birth weight and size of litter was based on 3353 cases, the other two correlations on 1317 cases.

In order to make a comparison of Minot's two alternatives,

these may be represented graphically in a single diagram.

Birth weight  $B$  is completely determined (in the mathematical sense rather than causally) by the prenatal growth curve and the age at which growth is interrupted by birth ( $G$ ). It is assumed that the rate of growth ( $R$ ) immediately before birth is a sufficient index of the growth function and that the rate of growth is uniform at this time to a sufficient degree of approximation. In substituting gestation period for interval a small correction is desirable. On grounds which need not be gone into here, it is estimated that the correlation between interval and true gestation period is about .95. No correction is necessary for birth weight since there is little or no growth in the first day after birth. The correlations involving interval must be divided by .95 to obtain estimates of those involving gestation period.

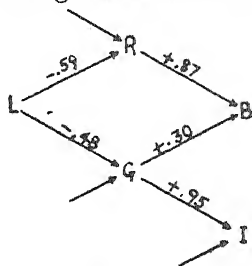


FIG. 12

Minot's problem resolves mathematically the analysis of the observed correlation between birth weight and size of litter into the sum of two composite path coefficients representing the two postulated paths of influence.

$r_{BG} = +.56$ ,  $r_{GL} = -.48$ , while  $r_{BL} = -.66$  is unchanged.

Minot's problem resolves mathematically the analysis of the observed correlation between birth weight and size of litter into the sum of two composite path coefficients representing the two postulated paths of influence.

$$r_{BL} = P_{BRL} + P_{BGL}$$

The method furnishes at once four equations for determining the values of the four path coefficients. One of these expresses the complete determination of  $B$  by  $R$  and  $G$ . The others are the expressions for the three known correlations.

$$(35) \quad P_{BR}^2 + P_{BG}^2 + 2 P_{BR} \cdot P_{BG} \cdot P_{RL} \cdot P_{GL} = 1$$

$$(36) \quad P_{BR} \cdot P_{RL} + P_{BG} \cdot P_{GL} = -.66$$

$$(37) \quad P_{BR} \cdot P_{RL} \cdot P_{GL} + P_{BG} = +.56$$

$$(38) \quad P_{GL} = -.48$$

These are not all linear equations, a condition which generally distinguishes this sort of application of the method from the calculation of partial regression coefficients. In the present case, however, there is no difficulty in the solution.

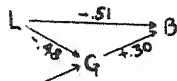
$$P_{BR} = +.87, \quad P_{BG} = +.30, \quad P_{RL} = -.59, \quad P_{GL} = -.48, \\ P_{BRL} = -.51, \quad P_{BGL} = -.15, \quad r_{BL} = -.66.$$

The result is an analysis of the correlation between birth weight and size of litter into two components whose magnitudes indicate that size of litter has more than three times as much linear effect on birth weight through the mediation of its effect on growth as through its effect on the length of the gestation period, contrary to the results of Minot's verbal analysis.

In this case, the answer to Minot's question might have been obtained from a set up mathematically identical with that used in multiple regression (after correcting the correlations with interval to obtain estimates of those with true gestation period.) By equation 24,

$$(39) \quad r_{BL} = P_{BL} + P_{BG} \cdot r_{GL}$$

$$(40) \quad r_{BG} = P_{BL} \cdot r_{GL} + P_{BG}$$



The term  $P_{BL} = -.51$  can be interpreted as measuring the influence of size of litter on birth weight in all other ways than through gestation period. In other cases, however, proper causal analysis may require a set up utterly different from that used in obtaining the best estimation equation. There is no routine method of making the proper diagram in the former case. This seems to have occasioned more misunderstanding than anything else among those who have attempted to apply the method. One author in a critique of the method, took the form of diagram intended to represent the sequential relations in the case of guinea pig weight and arranged some variables relating to basal metabolism in man in the same scheme in an arbitrary

FIG. 13

way and then complained of the meaningless and absurd results which he obtained!

### *Transpiration of Plants*

The contrast between the kind of set up appropriate to an estimation equation and that for evaluation of a causal interpretation was illustrated early (1921a) in connection with a study of the data of Briggs and Shantz on transpiration in plants. The reader is referred to the paper for the details, but it may be appropriate here to compare the different diagrams used. The authors obtained the total daily transpiration of a number of plants. The environmental factors studied were total solar radiation ( $R$ ), wind velocity ( $W$ ), air temperature in the shade ( $T$ ), rate of evaporation from a shallow tank ( $E$ ), and wet bulb depression, sheltered from sun, but not wind ( $B$ ). To avoid seasonal effects, the logarithms of ratios for successive days were used instead of absolute values.

An estimation equation for wet bulb depression was obtained in terms of wind velocity, solar radiation and temperature (Fig. 14).

It was pointed out that for causal analysis, radiation should be omitted as not affecting wet bulb depression in the shade, while a factor not directly measured, absolute humidity

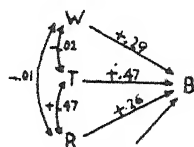


FIG. 14

( $H$ ) should be included. There should be complete determination of  $B$  by  $W$ ,  $T$  and  $H$ . As so arranged, there are two more unknown coefficients than known ones. It was assumed that there was no correlation between absolute humidity and wind velocity. The necessary additional equation was obtained from the theoretical multiple regression equation relating  $B$  to  $W$ ,  $T$  and  $H$ , by substituting the extreme differences in wet bulb depression, temperature and wind velocity of the average daily cycle and assuming the absence of any such cycle in absolute humidity. Possibly this was not wholly justified in this case. If so, no numerical evaluation of the chosen point of view could be made.

Even in such cases, the attempt at analysis by path coefficients may be valuable in locating deficiencies in the data already collected and suggesting the kinds of new data which should be obtained.

The final set up used in relating transpiration  $T_r$  and evaporation from a tank to wet bulb depression and the chosen environmental factors is given in figure 15 with the values of the path coefficients and correlations. Determinations were made for 10 varieties of plants. These gave fairly consistent results which are averaged in Fig. 15 although there were certain interesting differences. There was a marked difference between the transpiration of the plants and the rate of evaporation from the tank in the relative importance of the various factors.

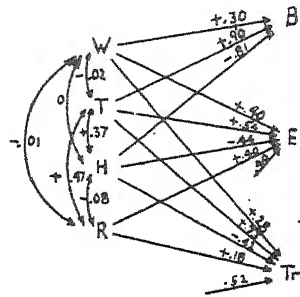


FIG. 15

### *The Relative Importance of Heredity and Environment*

Among the most satisfactory applications to causal relations are to problems of genetic determination. The development of an organism is the product of the confluence and interaction of two distinct streams of causation, heredity and environment. The interaction between the hereditary influences emanating from the nuclei of the cells of the organism and the influences coming from outside these cells, but largely from other parts of the body, where they in turn are the products of heredity and cell environment and so on back to the one cell stage are complex enough, but if we go back of this to the ultimate factors: the array of genes assembled at fertilization and the environmental conditions external to the organism, the sequential relations are for the most part clear. The problem is that of determining the relative importance of *differences* in heredity and of *differences* in environment in determining differences in the characteristics of individ-



graded at birth, but only very minor effects of this sort have been discovered experimentally, the most important (contributing .036 to the correlation of litter mates) being an effect of the age of the mother. The high degree of asymmetry of the pattern in individual animals is in harmony with a large element of chance (somatic mutation?) in the determination of pigmented areas.

The above estimates ( $h^2 = .38$ ,  $d^2 = .53$ ,  $e^2 = .09$ ) are estimates of the portion of the variance due to heredity, non-genetic factors peculiar to individuals, and common environment, respectively. They are the portion of the variance which should be eliminated by control of each factor. It is not possible to control the rather intangible environmental factors but hereditary *variation* can be eliminated by close inbreeding (decrease of heterozygosis being about 19% per generation under brother-sister mating). It happened that a number of piebald stocks were on hand, each descended from a single mating after several generations of inbreeding. These differed markedly in average percentage of white in the coat, although individuals of each varied widely about their family averages. Crosses between strains at opposite extremes gave intermediate offspring, justifying the assumption of no dominance. The family (No. 35) most advanced in inbreeding was descended from a single mating in the 12th generation of brother-sister mating, but even in it there was variation from nearly solid color to solid white. As expected by theory, very little, if any, of this variability was hereditary. The correlation between parent and offspring was only  $+0.024 \pm .020$ . The correlation between litter mates was  $+0.103 \pm .025$ , again indicating only a small amount of influence of environment common to litter mates.

The standard deviation, measured on an appropriate scale<sup>4</sup>

<sup>4</sup> On a percentage scale of measurement, necessarily limited at 0% and 100%, a given factor has more effect near the middle of the range than near the limits. The appropriate transformation of the scale  $X$ , ranging from 0 to 1 is  $X' = \text{Prf}(x - 50)$ , where  $\text{Prf}$  is the inverse probability function,

the direct function being defined in the form  $\text{Prf } x' = \frac{1}{\sqrt{2\pi}} \int_0^{x' - \frac{x'}{2}} e^{-\frac{z^2}{2}} dz$ .  
(Wright 1926a).



came out .574 (about 22% of the area of coat in the neighborhood of 50%). In the random bred stock the standard deviation was 0.782 (about 28%). The variance of the stock in which hereditary variation had been eliminated was thus  $54\% = \left(\frac{.574}{.782}\right)^2$

of that of the random bred stock. This agrees as well as could be expected with the estimate of 62% of the variance of the latter as nongenetic, based on the parent offspring correlation, although not as well as an earlier estimate made when the numbers were smaller (nongenetic variance 58% as deduced for a parent-offspring correlation of +.21 in random stock, variance of inbred family 57% of that of random stock, i.e.  $\frac{.364}{.643}$  ).

### Case of Human Intelligence

Another illustration of the difference between a quantitative interpretation and a multiple regression formula has been given (Wright 1931a) using data of Miss B. S. Burks, on the roles of heredity and environment in determining human intelligence. These data consisted of intelligence tests of 104 California children, tests of their parents and in addition grades of home environment. Similar data were obtained of 206 children adopted at an average age of 3 months, and of their foster parents and home environments. The correlations as used were corrected by Miss Burks for attenuation.

If the purpose is to obtain the best estimation for children in terms of their parents and environments, the variables are to be related as in figure 17 in which  $C$  is child's intelligence,  $P$  is midparent and  $E$  is the measure of home environment.

#### Normal Equations

#### Children

	(Own)	(Adopted)
(43) $P_{CE} \cdot r_{EP} + P_{CP} = r_{CP} = +.61$		+ .23
(44) $P_{CE} + P_{CP} \cdot r_{EP} = r_{CE} = +.49$		+ .29
(45) $r_{EP} = +.86$		+ .86
Solutions:		
$P_{CP} = +.72$		-.07
$P_{CE} = -.13$		+ .35



FIG. 17

The solutions of the normal equations in the two bodies of data give what at first sight appear to be contradictory results. There is no apparent reason why environment should not play as great a role in shaping intelligence in one case as in the other, yet it turns out that while the partial regression of child's IQ on home environment is significantly positive in the foster data, it is negative as far as it goes, in the case of own children.

The point that is sometimes overlooked is that the arrangement for obtaining the best possible prediction equation does not necessarily yield coefficients which have any simple interpretation. This is obviously the case here. If child's IQ is affected both by heredity and environment, the same is presumably true of parental IQ. In so far as the latter is determined by environment it is not a causal factor in relation to child's *heredity*. A diagram intended to represent causal relation *must* represent parental IQ as merely correlated (two headed arrows) with child's heredity and child's environment. Another complication which must be represented is the correlation of heredity with environment. Good heredity in a family will tend to create a good environment and vice versa. The simplest possible *interpretative* diagram for own children is thus of the type of figure 18. That for foster children is given in figure 19.

Even these are doubtless too simple since heredity is represented as the only factor apart from the measured environment. Any estimates of the importance of hereditary variation will thus be maximum.

The two correlations given by Miss Burks in the case of the foster data (Fig. 19)

$$(r_{CE} = +.29, \quad r_{CP} = .23)$$

yield the value  $r_{EP} = +.79$

for the correlation between home environment and midparental IQ. The actual correlation was

not published for the foster data, but there is no reason why it

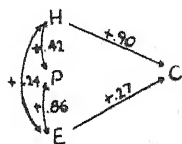


FIG. 18

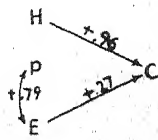


FIG. 19

should differ significantly from that in the other data in which the value was  $+.86$ . There is reasonable agreement.

In the case of own children, three correlations of interest here, were published,  $r_{CE} = +.49$ ,  $r_{CP} = +.61$ ,  $r_{EP} = +.86$ . But this is not enough to give a solution for the coefficients of the 5 indicated paths. The assumption of complete determination of  $C$  by  $H$  and  $E$  gives a fourth equation, still an inadequate number. No solution is possible, a situation which as previously noted, very frequently arises in such analysis, even when one makes the most simplified possible qualitative representation of the causal relations. A great deal of utterly unwarranted verbal interpretation of correlation coefficients would be avoided if the authors took the trouble to represent their ideas in diagrammatic form and noted whether or not the number of equations possible from the data (known correlation coefficients and known cases of complete determination) was as great as the number of paths in this diagram.

In the present case, another equation can be obtained by borrowing from the foster data. Environment should make approximately the same contribution to IQ in both groups of children. The concrete partial regression coefficients

$$C_{CH} \left( = \bar{P}_{CH} \frac{\sigma_C}{\sigma_H} \right) \quad \text{and} \quad C_{CE} \left( = \bar{P}_{CE} \frac{\sigma_C}{\sigma_E} \right)$$

should thus be approximately the same in the foster as in the own children. Assuming that  $\sigma_H$  and  $\sigma_E$  are the same in both cases, the ratio  $\frac{\bar{P}_{CE}}{\bar{P}_{CH}}$  from the foster data may be accepted for the group of own children. The five equations now available are as follows:

	Equations	Solution
(46)	$r_{EP} = +.86$	
(47)	$r_{CE} = +.49 = \bar{P}_{CE} + \bar{P}_{CH} \cdot r_{HE}$	$\bar{P}_{CE} = +.27$
(48)	$r_{CP} = +.61 = \bar{P}_{CE} \cdot r_{EP} + \bar{P}_{CH} \cdot r_{HP}$	$\bar{P}_{CH} = +.90$
(49)	$\bar{P}_{CE} = +.302 \bar{P}_{CH}$	$r_{HP} = +.42$
(50)	$\bar{P}_{CE}^2 + \bar{P}_{CH}^2 + 2 \bar{P}_{CE} \cdot \bar{P}_{CH} \cdot r_{HE} = 1$	$r_{HE} = +.24$

The solution assigns reasonable values in all cases and shows that there was no real disagreement involved in the relation of the two groups of children to their environments.

It was noted that this analysis gives a maximum estimation of the role of heredity. An attempt was made to obtain a minimum estimate compatible with acceptance of the observed correlations, by carrying the analysis back a generation and assuming as much similarity in the determining factors of successive generations as the data permit.

Such analysis requires separate treatment of heredity ( $H$ ) as a factor of development, and heredity or genotype ( $G$ ) as the linear system of gene effects which best approximates the former. Departures from linearity in the effects of allelomorphs (dominance) and in the effects of nonallelomorphs (epistasis) are common. Moreover there may be non-linearity in the combination effects of heredity and environment. Thus a certain genetic complex in the guinea pig ( $c^d c^a BB$ ) produces more melanin pigment at low temperatures than does a certain other ( $CC\&\&$ ) but less at high temperatures (Wright 1927). The subject is too involved for detailed discussion here but it may be noted that in general correlations between deviations due to dominance and epistasis must be taken account of.

In the case of Miss Burk's data, there is no possible way of distinguishing the effects of environmental factors not included in the measurement of home environment from the contributions of dominance and epistasis or from non-linearity in the combination effects of heredity and environment. In the attempt at obtaining a minimum estimate of heredity, these three very diverse factors were put together in a miscellaneous group  $M$ . The diagram of relation used is given in Fig. 20. Child's genotype ( $G$ ) is represented as partially determined by midparental genotype ( $G'$ ), the residual variability being that of Mendel's

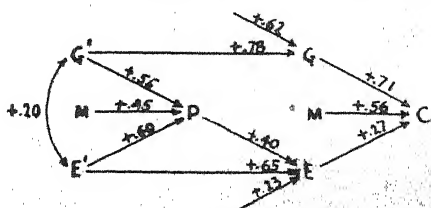


FIG. 20

lian segregation. Child's environment is treated as in part determined directly by midparental intelligence  $P$  and in part tracing to the environment of the preceding generation,  $E'$ .

The path coefficient relating genotype of midparent to that of child could be estimated, assuming Mendelian heredity and taking into account a correlation of  $+ .70$  between father and mother. It turned out to be mathematically impossible to assign the same values to the path coefficients of the parental generation as in the offspring generation, but this is not surprising since the parents were tested as adults instead of young children. The solution for the parent generation was to some extent indeterminate but within rather narrow limits, on making what seemed the most reasonable assumptions. The values reached are given in figure 20. The path coefficient for influence of hereditary variation lies between the limits  $+ .71$  (if dominance and epistasis are lacking) and  $+ .90$ .

#### *Analysis of Size Factors*

The first published application of the method was to the interpretation of a system of correlations of bone measurements (length and breadth of skull, lengths of humerus, femur and tibia) in a population of rabbits (Wright 1918). The 10 observed correlations were accounted for primarily as due to a single general factor (not necessarily acting proportionately on the 5 variates). The residuals which appeared were attributed to group factors.

In a recent paper (1932b) the same figures, two other sets of figures for rabbit populations ( $F_1$  and  $F_2$  of a wide cross) and figures from a flock of hens have been analyzed by a somewhat improved method. A set of  $n$  variables yields  $\frac{n(n-1)}{2}$  correlation coefficients and hence the same number of observation equations of the type  $r_{AB} = P_{AG} \cdot P_{BG}$ , where  $A$  and  $B$  are two of the variables and  $G$  is the general factor and it is assumed for the moment that the correlations are due solely to differences in general size. The residuals are minimized by the method of least squares.

This method necessarily gives residuals which are as likely to be negative as positive. The interpretation is more satisfactory if the path coefficients relating each measurement to the general factor are all reduced by the proportion necessary to eliminate significant negative residuals. It happened that in each of the 4 sets of data studied, the most important negative residuals were those between the skull and hind leg measurements, and the method followed was to eliminate the average of these.

The important positive residuals in all cases indicated natural group factors—a head group, a general leg group, a foreleg group, a hind leg group (in the one case in which both humerus and ulna were measured) and a hind leg group. Other indications such as a slightly closer relation between head and foreleg than between head and hind leg, slightly closer relation between proximal leg bones (humerus and femur) than between non-homologous

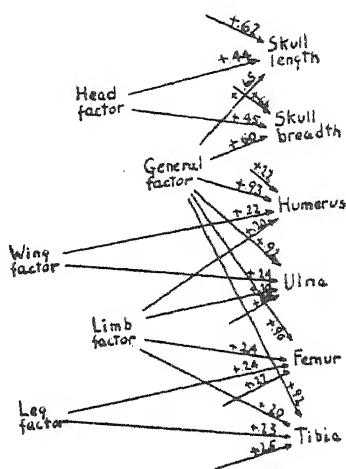


FIG. 21

and hind leg bones (humerus and tibia) were less certain. Figure 21 shows the system of path coefficients arrived at in the case of the fowl measurements. The squares of these give the degree of determination in each case by the general factor, the group factors and special factors.

### *The Use of Partial Correlation in Interpretation*

Partial correlation coefficients have sometimes been used in the attempt to interpret systems of correlated variables apparently on the theory that the reduction or elimination of a correlation between two variables on holding a third constant demonstrates the latter to be causally responsible for the correlation. The

method at first sight seems analogous to that of the experimentalist in attempting to control all sources of variation except those in which he is interested. This, however, is a delusion in the case of correlation (as opposed to regression) coefficients (Wright 1921a) and the method of path coefficients was developed because of the unsatisfactory nature of interpretation based on partial correlation. As R. A. Fisher (1925) has stated, "In no case, however, can we judge whether or not it is profitable to eliminate a certain variable unless we know or are willing to assume a qualitative scheme of causation."

This point can be illustrated by considering a system of 3 variables,  $A$ ,  $B$  and  $C$  in which the following correlations have been found.

$$r_{AB} = .50 \quad r_{BC} = .50 \quad r_{AC} = .25.$$

By substitution in the usual formula,  $r_{AC.B} = 0$ . This is compatible with the interpretation, represented in figure 22, that  $B$  is an intermediary in a single chain of causation connecting  $C$  and  $A$ .

$$r_{AC} = P_{AB} \cdot P_{BC} = r_{AB} \cdot r_{BC} = .25.$$

Another interpretation is that  $B$  is the only common factor

$$r_{AC} = P_{AB} \cdot P_{CB} = r_{AB} \cdot r_{BC} = .25.$$

But it is also possible that  $B$  may be the product of the interaction of two correlated factors  $A$  and  $C$

$$r_{AB} = P_{BA} + P_{BC} \cdot r_{AC} = .50.$$

Finally  $A$ ,  $B$  and  $C$  may be correlated with each other through reciprocal interactions or through complexes of unknown common factors, making impossible anything beyond the mere descriptive use of the correlation coefficients, or the calculation of estimation equations.

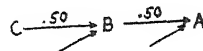


FIG. 22

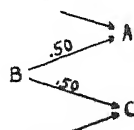


FIG. 23

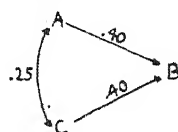


FIG. 24

The first step in the application of the method of path coefficients is to bring clearly into the open the system of functional relations among the variables which seems significant for purposes of interpretation. In the majority of cases, verbal interpretations which seem reasonable enough as long as the basic postulates

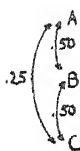


FIG. 25

are kept discretely in the subconscious mind become obviously crude and inadequate when expressed in a diagram. Occasionally, however, statistical systems are capable of some interpretation.

### *Difficulties in Causal Analysis*

There are a great many systems of correlated variables for which no interpretation can be suggested in terms of sequential relations. Among these are cases in which there is prevaillingly mutual interaction between the variables instead of action in one direction. The branches of science differ considerably in the type of relation which predominates.

As already noted the developmental process of organisms is essentially a one way process, and the ultimate factors of development, heredity and environment act on it without being acted upon. A method of analysis which takes account of the sequential relations is thus imperatively called for in genetics.

A case in which such analysis would not be possible may be illustrated by the relations among the various properties of the blood, as discussed by L. J. Henderson. The physiological mechanisms are such that alteration of any one brings about immediate readjustments in the values of the others. What one wishes to determine are the functional relations, whether in the form of equations or of nomograms. If such a system were studied by correlational methods the best that could be done would be to attempt to approximate the functional relations by multiple regression (linear or curvilinear as the case required).

There is usually rapid reciprocal action among the variables of interest to the economist or sociologist and the correlations



among the simultaneous deviations cannot, in most cases, be treated as due to lines of one way causation among these variables themselves. Thus the price of a commodity cannot properly be treated as caused by the amount marketed or vice versa. The exception is where one variable is clearly external to the social system in question as is the influence of weather on crop yield.

There is more likelihood of being able to represent the various simultaneous deviations as direct consequences of the system of deviations of the preceding year (together with the clearly external contemporary factors) but even here, a causal diagram can be set up only after a most careful consideration of the realities of the case. There may be lags of greater duration than one year and a correlation between two variables in successive years may trace to more remote common factors rather than to a direct line of causation from the earlier to the later.

#### *Corn and Hog Correlations*

These points were illustrated by a study of corn and hog correlations (Wright 1924). An attempt was made to analyze the play of interacting factors responsible for the annual fluctuations from the general trends in production and price of hogs during the relatively undisturbed period between the Civil War and the World War. It was shown that variation in the corn crop and certain interrelations among the hog variables themselves determined from 75 to 85% of the variance of the latter. The annual fluctuations about the trend during the period of years from 1871 to 1915 inclusive (so far as data were available) were found for corn acreage, yield, crop and price and for western and eastern wholesale hog packs and for farm price of hogs. The fluctuations were found separately for the summer and winter seasons for western wholesale hog pack and the corresponding live weight, pork production (product of preceding) and price. Correlation coefficients were found not only for the same year but between variables separated by one, two and often three years. Altogether 510 correlation coefficients were calculated.

Most of these coefficients could be given reasonable enough verbal interpretations, but there was no assurance that the "obvious" interpretation in one case, was compatible with an equally "obvious" interpretation in another. The problem was to represent all of these verbal interpretations in a single diagram and determine path coefficients which would account simultaneously for the entire system of correlation coefficients. With 510 correlation coefficients and 4 cases of complete determination, one could write 514 simultaneous equations to determine the values of whatever system of path coefficients had been used. Theoretically one could introduce the same number of different paths into the diagram. It would not be practicable, however, to deal with such a large number of unknown quantities and even if practicable, the complexity of the system would defeat the purpose of the analysis. The problem thus resolved into the discovery of a simple system of relations which would give a reasonably close approximation to all of the correlation coefficients.

It has been emphasized that the method of path coefficients is not intended to accomplish the impossible task of deducing causal relations from the values of the correlation coefficients. It is intended to combine the quantitative information given by the correlations with such a qualitative information as may be at hand on causal relations to give a quantitative interpretation. The analysis of cases such as the present and that preceding (size factors), in which the equations far outnumber the coefficients to be determined, may appear to be exceptions to this statement, but even here only such paths are tried which are appropriate in direction in time and which can be given a rational interpretation.

Considerable experimentation was necessary before a simple system could be found which gave even moderately satisfactory results. The procedure followed was to list the highest five correlations of each variable with a preceding variable. It turned out that the corn variables were so nearly independent of conditions in preceding years that they might be treated practically as

independent in relation to the hog situation. The variations in corn crop depended largely on variations in yield ( $P_{CY} = +.90$ ) and secondarily on variations in acreage ( $P_{CA} = +.45$ ). Corn price showed a correlation of  $-.80$  with the crop.

Among the hog variables, the maximum correlations were with those which indicated most directly the amount of breeding (average summer weight ( $sw$ ) of the same year, winter pack ( $wP$ ) a year and a half later, between which there was a correlation of  $+.78$ ), and with the preceding prices of corn and of hogs. The four variables: breeding ( $B$ ), summer ( $S$ ) and winter ( $w$ ) price of hogs and price of corn ( $P$ ) were thus chosen as a central system. 36 equations could be written involving these (using jointly the two indicators of breeding). Values of 13 path coefficients were tested by repeated trial and error until it seemed that no change (of the order of  $.05$ ) would give improvement. The system reached is shown in figure 26 in which primes refer to preceding years.

The other variables were then appended to this system, also by the trial and error method. Corn crop was used in place of corn price, however. The results are shown in figures 27 and 28. These bring out the very different characteristics of the summer pack ( $SP$ ) (consisting of a very heterogenous lot of hogs) and the winter pack, ( $wP$ ) largely consisting of the spring pig crop. Average summer and winter live weights are represented by ( $sw$ ) and ( $ww$ ) in figure 27.

The general conclusions were that the dominating features of the hog situation are the corn crop and its price, and an innate tendency to fall into a cycle of successive overproduction and underproduction, two years from one extreme to the other, depending mainly on two compound paths:

$$P_{BWB''} = -.42 \text{ and } P_{BSB''} = -.14.$$

The 32 indicated path coefficients together with 10 others relating total annual western pack to its components, eastern pack to western pack and prices, and farm price to packer's price, ac-

counted for the 510 observed correlation coefficients with an average error of only .09 neglecting sign. The most serious discrepancies were in certain correlations involving corn acreage and yield which were intentionally ignored for the sake of avoiding complexity in the relations of the more important variables.

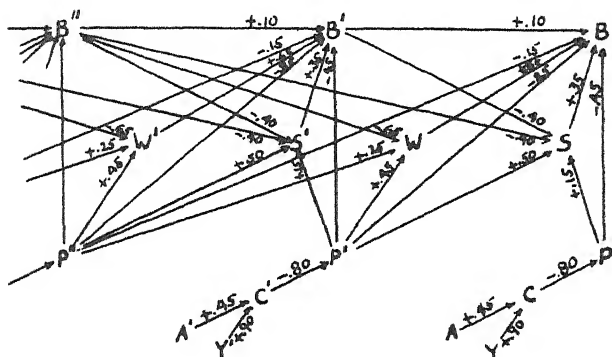


FIG. 26

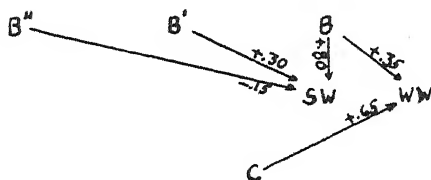


FIG. 27

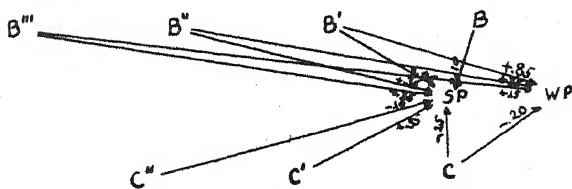


FIG. 28

### *The Elasticities of Supply and Demand*

In the preceding illustration, market supplies, prices, etc. were related to preceding conditions largely by a trial and error process of finding the system which would work best and without much

regard for theoretical considerations. In the following more theoretical approach I have collaborated with Dr. P. G. Wright. The purpose is to interpret observed series of prices and quantities marketed as functions of two hypothetical variables, the conditions of supply and demand. Only a brief reference has previously been published (P. G. Wright 1928).

The demand for a given commodity and given market is treated as that function of all economic factors (prices, wages, etc.) which determines the quantity which would be purchased under any set of postulated conditions. The supply function, similarly, is treated as that function of all economic factors (prices, manufacturing costs, weather, etc.) which determines the quantity which would be offered for sale under any set of postulated conditions. The actual values which these functions take at a given moment tend to be the same, the price of the commodity itself being the immediate factor which shifts to such a value as to make them identical.

We shall deal with the annual percentage deviations in quantities and prices, whether from the preceding year or from the estimated trend of a series of years, instead of absolute values. The relative merits of these two procedures need not be gone into.

Let  $X$  represent values on a scale of *percentage* change in quantity and  $Y$  values on a scale of *percentage* change in the price of the commodity in question. Let  $Z_1, Z_2$ , etc. represent other economic factors of demand or supply or both on whatever scales are most suitable. The demand and supply functions themselves as percentage deviations in quantities under postulated conditions may be represented by  $X_d$  and  $X_s$  respectively.<sup>5</sup>

$$(51) \quad X_d = f_d (Y, Z_1, Z_2, \dots, Z_d, \dots)$$

$$(52) \quad X_s = f_s (Y, Z_1, Z_2, \dots, Z_d, \dots)$$

<sup>5</sup> If the absolute quantities are represented by  $U$  and the absolute prices by  $V$ ,  $X = \frac{\Delta U}{U}$  and  $Y = \frac{\Delta V}{V}$ . It is customary to define the demand and supply functions in terms of the absolute values, but for the present purpose it is more convenient to define them in relation to the percentage deviation.

Assume that these functions are of such a nature that the deviations in price can be separated linearly from the other factors to a sufficient degree of approximation. This does not imply lack of correlation between price and the others.

$$(53) \quad X_d = \eta Y + D \quad \text{where} \quad D = f'_d(Z_1, Z_2, \dots, Z_d, \dots)$$

$$(54) \quad X_s = \epsilon Y + S \quad \text{where} \quad S = f'_s(Z_1, Z_2, \dots, Z_s, \dots)$$

The demand function is here analyzed into two variable components, a multiple of the price deviation ( $\eta Y$ ) and the deviation ( $D$ ) in the quantity which would be purchased if there were *no price deviation* ( $Y=0$ ). The supply function is similarly analyzed into a different multiple of the price deviation ( $\epsilon Y$ ) and the deviation ( $S$ ) in the quantity which would be offered for sale in the absence of a price deviation. Thus  $D$  and  $S$  measure the strength of demand and supply apart from price and will be spoken of as measures of demand and supply.

For given values of  $D$  and  $S$  the equations define two straight lines which describe the momentary demand and supply situations respectively (Fig. 29). Their slopes relative to the  $Y$ -axis are given by  $\eta$  and  $\epsilon$  respectively. These slopes are in accordance with the customary definitions of the elasticities of demand and supply, recalling that  $X$  and  $Y$  are percentage deviations.<sup>6</sup>

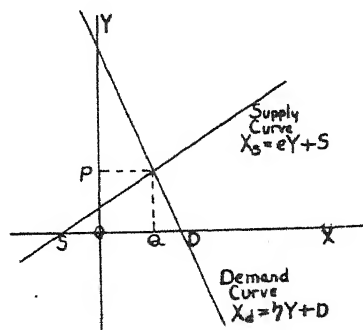


FIG. 29

According to the usual theory, the actual quantity which changes hands and the actual price are determined by the point of intersection of the supply and demand curves. Under the approximations previously assumed, and assuming constancy of the elasticities, but variation of  $D$  and  $S$ , the percentage deviations

<sup>6</sup> The ratio  $\frac{X}{Y} = \frac{\Delta u / u}{\Delta v / v}$  where  $u$  and  $v$  are absolute quantities and prices respectively. The ratio  $\frac{X_s}{Y}$  is the elasticity of supply ( $\epsilon$ ) if  $S=0$ , and  $\frac{X_d}{Y}$  is the elasticity of demand ( $\eta$ ) if  $D=0$ .

in quantity ( $Q$ ) and in price ( $P$ ) are linear functions of  $D$  and  $S$ . Their values may be represented as determined by multiple regression equations. It will be convenient to use single letters for the path coefficients.

$$(55) \quad P = p_1 \frac{\sigma_P}{\sigma_D} D + p_2 \frac{\sigma_P}{\sigma_S} S$$

$$(56) \quad Q = q_1 \frac{\sigma_Q}{\sigma_D} D + q_2 \frac{\sigma_Q}{\sigma_S} S$$

The elasticity of supply may be obtained from the ratio of  $Q$  to  $P$  under a fixed average supply situation ( $S=0$ ) but variable demand.

$$(57) \quad e = \frac{q_1 \sigma_Q}{p_1 \sigma_P}$$

Similarly, elasticity of demand is given by the ratio of  $Q$  to  $P$  when  $D$  equals zero.

$$(58) \quad \eta = \frac{q_2 \sigma_Q}{p_2 \sigma_P}$$

Since the standard deviations are obtainable directly from the data it is merely necessary to find the values of the path coefficients in order to calculate the two elasticities.

A diagram can be set up as in Fig. 30 indicating primarily that  $P$  and  $Q$  are different linear functions of  $D$  and  $S$ . Three equations can be written at once; two indicating complete determination of  $P$  and  $Q$  by  $D$  and  $S$ , and one representing the correlation between  $P$  and  $Q$ .

$$(59) \quad p_1^2 + p_2^2 + 2 p_1 p_2 r_{SD} = 1$$

$$(60) \quad q_1^2 + q_2^2 + 2 q_1 q_2 r_{SD} = 1$$

$$(61) \quad p_1 q_1 + p_2 q_2 + (p_1 q_2 + p_2 q_1) r_{SD} = r_{QP}$$

Unfortunately these three equations involve 5 unknowns. Other data must be brought to bear on the problem before any solution is possible.

The diagram suggests two possible sources of

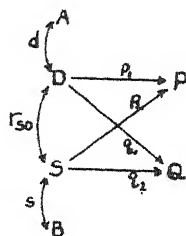


FIG. 30

additional data. If any measurable quantity ( $A$ ) can be found which is correlated with the demand situation but which can safely be assumed to be independent of the supply situation, ( $r_{AS} = 0$ ) we can write two new equations representing the correlations  $r_{AP}$  and  $r_{AQ}$  respectively at the expense of only one additional unknown ( $r_{AD} = d$ ). We have now 5 equations and 6 unknowns. If it can safely be assumed that there is no correlation between the demand and supply situations ( $r_{SD} = 0$ ), a solution is possible. If such an assumption with regard to  $r_{SD}$  does not seem justified, it may be possible to find a quantity ( $B$ ) correlated with the supply situation (as measured by  $S$  but of such a nature that no correlation with the demand situation need be postulated. The correlation  $r_{BP}$  and  $r_{BQ}$  make possible two more equations, with only one more unknown ( $r_{SB} = s$ ) bringing the number of equations and unknowns both up to 7. The path coefficients and hence the elasticities are now determinate. The additional equations are as follows:

$$(62) \quad r_{AP} = p_1 d \qquad (64) \quad r_{BP} = p_2 s$$

$$(63) \quad r_{AQ} = q_1 d \qquad (65) \quad r_{BQ} = q_2 s$$

The hog and corn data referred to in the preceding section were not obtained with the present purpose in mind, but may furnish rough illustrations of the method. The total weight of hogs, marketed at the principal markets in the summer season (March to October) 1889-1914, and the reported price may be considered first. Absolute instead of percentage deviations from trend were used but the correlations should not be affected much and coefficients of variation may be used in place of the standard deviations on a percentage scale. The most important single factor affecting the summer hog pack was shown to be the corn crop of the preceding year. It is assumed that it is a factor of type  $B$ , correlated with the supply situation as measured by  $S$  but not with the demand for pork as measured by  $D$ . It is further as-



sumed that there was no correlation between the supply and demand situations.

### Data

Coefficient of variation—Price	$\sigma_p = 15.86$
Quantity	$\sigma_q = 10.89$
Correlation—Price with quantity	$r_{pq} = -.63$
Correlation—Hog price with preceding corn crop	$r_{pB} = -.47$
Correlation—Weight of pack with preceding corn crop	$r_{qB} = +.64$

Equations	Solution	
$p_1^2 + p_2^2 = 1$	$p_1 = +.686$	$e = +.133$
$q_1^2 + q_2^2 = 1$	$p_2 = -.728$	$\eta = -.944$
$p_1 q_1 + p_2 q_2 = -.63$	$q_1 = +.132$	
$p_2 s = -.47$	$q_2 = +.991$	
$q_2 s = +.64$	$s = +.646$	

The solution indicates very little elasticity of supply ( $e = +.133$ ) but a very considerable elasticity of demand ( $\eta = -.944$ ).

Similar data were given for the winter weight of pack (1870-1914). The largest correlation with a factor of preceding years was with average summer live weight of hogs, one and one half years before. This factor (an index of amount of breeding) is again assumed to be related to the supply but not to the demand situation, and again it is assumed that the supply and demand situations vary independently of each other.

Data	Solution	
$\sigma_p = 12.59$	$p_1 = +.656$	$e = +.110$
$\sigma_q = 18.75$	$p_2 = -.755$	$\eta = -.884$
$r_{pq} = -.68$	$q_1 = +.108$	
$r_{pB} = -.63$	$q_2 = +.994$	
$r_{qB} = +.83$	$s = +.835$	

The results are remarkably close to those of the quantity and price of summer pork. The low elasticities of supply are to be expected of an agricultural commodity the quantity of which is largely determined in advance and by factors independent of the market demand and which once produced must largely be marketed.

I am indebted to my colleague, Professor Henry Schultz, for data on the quantity and price of potatoes marketed annually from 1896 to 1914 and the suggestion that it would be interesting material for analysis by this method. Trends had been fitted by Professor Schultz and trend ratios of quantity and price obtained.

#### Data

Standard deviation of price ratios  $\sigma_p = .185$

Standard deviation of quantity ratios  $\sigma_q = .130$

#### Correlations

Price—quantity (same year)  $r_{pq} = -.852$

Price—quantity (preceding year)  $r_{p'q'} = +.570$

Price—price (preceding year)  $r_{pp'} = -.562$

Quantity—quantity (preceding year)  $r_{qq'} = -.522$

Quantity—price (preceding year)  $r_{q'p'} = +.651$

It is assumed again that there is no correlation between supply and demands situations ( $r_{sp} = 0$ ) and that the price (as a trend ratio) is a factor of type *B* affecting the supply of the following year but without influence on the demand of the following year. The solution is as follows:

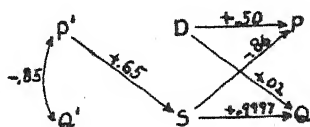


FIG. 31

$$p_1 = +.503 \quad g_1 = +.024 \quad e = +.034$$

$$p_2 = -.864 \quad g_2 = +.9997 \quad \eta = -.815$$

$$s = +.651$$

Figure 31 gives a graphical representation of the relation.

Again the virtual absence of elasticity of supply might perhaps have been anticipated. The size of crop is largely determined

before the price is known and the crop must be disposed of regardless of price. It is to be noted, however, that this result came out quite independently of any such assumption.<sup>7</sup> There are other checks on the theory. Two of the correlations reported above have not been used. According to the diagram of relations

$\kappa_{QA'} = g_2 s \kappa_{PA} = -.554$ . The observed value,  $-.522$ , is in good agreement. Also  $\kappa_{PA'} = p_2 s \kappa_{PA} = +.479$ .

The agreement with the observed value of  $+.570$  is not as good as in the previous case, but considering the small number of years, is not bad.

The absence of elasticity of supply in the case of potatoes applies only within a single year. The fact that the supply is strongly correlated with the price of the preceding year  $+.651$  indicates that in the long run there is considerable elasticity. The method of path coefficients readily lends itself to deduction of this long time elasticity.

Let  $\bar{P}$ ,  $\bar{Q}$ ,  $\bar{A}$  and  $\bar{B}$  be the hypothetical averages of  $P$ ,  $Q$ ,  $A$  and  $B$  respectively over an indefinite ( $n$ ) period of years. The problem is to deduce the elasticities toward which the long time supply and demand curves tend, from knowledge merely of the correlations from year to year. The following equation can be written from figure 32, where  $a$ ,  $b$ ,  $c$  and

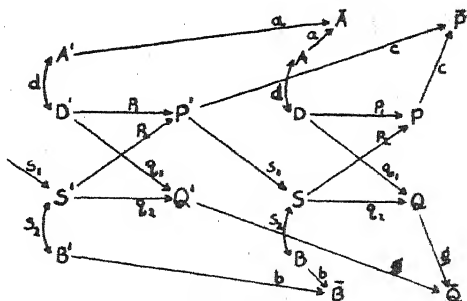


FIG. 32

$g$  are path coefficients pertaining to the paths indicated.

<sup>7</sup> In two other cases studied by this method (P. G. Wright 1928) very different results were obtained. In the case of butter, the elasticity of supply came out 1.43, of demand  $-.62$ . In the case of flax seed, the elasticity of supply came out even greater, 2.39, while that of demand was  $-.80$ . But these are cases in which a high elasticity of supply is to be expected on a priori grounds. It is interesting to note that in cases in which it seems justifiable to assume a priori that there is no elasticity of supply ( $e=0$ ), it follows that  $g_1=0$ ,  $g_2=1$ ,  $p_1=\kappa_{QP}$  (still assuming  $\kappa_{SD}=0$ ) and finally that  $\eta = \frac{\sigma_a}{\kappa_{PQ} \sigma_P} = \frac{1}{\rho_{PQ}}$ .

$$\begin{aligned}
 (66) \quad \eta_{\bar{P}\bar{A}} &= \pi c \eta_{P\bar{A}} = \pi c (p_1 \eta_{D\bar{A}} + p_2 \eta_{S\bar{A}}) \\
 &= \pi c (p_1 \eta_{D\bar{A}} + p_2 s_1 \eta_{P\bar{A}}) = \frac{\pi c p_1 \eta_{D\bar{A}}}{1 - p_2 s_1}
 \end{aligned}$$

$$\begin{aligned}
 (67) \quad \eta_{\bar{P}\bar{B}} &= \pi c \eta_{P\bar{B}} = \pi c (p_2 \eta_{S\bar{B}}) \\
 &= \pi c p_2 (s_1 \eta_{P\bar{B}} + s_2 t) = \frac{\pi c p_2 s_2 t}{1 - p_2 s_1}
 \end{aligned}$$

$$\begin{aligned}
 (68) \quad \eta_{\bar{Q}\bar{A}} &= \pi g \eta_{Q\bar{A}} = \pi g (q_1 \eta_{D\bar{A}} + q_2 \eta_{S\bar{A}}) = \pi g (q_1 \eta_{D\bar{A}} + q_2 s_1 \eta_{P\bar{A}}) \\
 &= \pi g (q_1 \eta_{D\bar{A}} + \frac{q_2 s_1 p_1 \eta_{D\bar{A}}}{1 - p_2 s_1}) = \pi g \eta_{D\bar{A}} \left( \frac{q_1 + q_2 s_1 p_1 - q_1 p_2 s_1}{1 - p_2 s_1} \right)
 \end{aligned}$$

$$\begin{aligned}
 (69) \quad \eta_{\bar{Q}\bar{B}} &= \pi g \eta_{Q\bar{B}} = \pi g q_2 \eta_{S\bar{B}} = \pi g q_2 (s_1 \eta_{P\bar{B}} + s_2 t) \\
 &= \pi g q_2 (s_1 p_2 \eta_{S\bar{B}} + s_2 t) = \frac{\pi g q_2 s_2 t}{1 - p_2 s_1}
 \end{aligned}$$

$$(70) \quad \bar{p} = \frac{\Sigma p}{n} = c \frac{\sigma_p}{\sigma_p} \leq p \quad \therefore \bar{\sigma}_p = \frac{\sigma_p}{nc}$$

$$(71) \quad \bar{q} = \frac{\Sigma q}{n} = g \frac{\sigma_q}{\sigma_q} \leq q \quad \therefore \bar{\sigma}_q = \frac{\sigma_q}{ng}$$

Let  $\eta_L$  and  $e_L$  be the elasticities of long time demand and supply respectively.

$$\begin{aligned}
 (72) \quad \eta_L &= \frac{\eta_{\bar{Q}\bar{B}} \bar{\sigma}_q}{\eta_{\bar{P}\bar{B}} \sigma_p} = \left( \frac{\pi g q_2 s_2 t}{1 - p_2 s_1} \right) \left( \frac{1 - p_2 s_1}{\pi c p_2 s_2 t} \right) \frac{\bar{\sigma}_q \pi c}{ng \sigma_p} \\
 &= \frac{q_2 \sigma_q}{p_2 \sigma_p}
 \end{aligned}$$

$$\begin{aligned}
 (73) \quad e_L &= \frac{\eta_{\bar{Q}\bar{A}} \bar{\sigma}_q}{\eta_{\bar{P}\bar{A}} \sigma_p} = \pi g \eta_{D\bar{A}} \cdot \frac{q_1 + q_2 s_1 p_1 - q_1 p_2 s_1}{1 - p_2 s_1} \cdot \frac{1 - p_2 s_1}{\pi c p_1 \eta_{D\bar{A}} \pi g \sigma_p} \bar{\sigma}_q \\
 &= \frac{q_1 + q_2 s_1 p_1 - q_1 p_2 s_1}{p_1} \cdot \frac{\bar{\sigma}_q}{\sigma_p} = e + s_1 \left( q_2 - \frac{q_1 p_2}{p_1} \right) \frac{\bar{\sigma}_q}{\sigma_p}
 \end{aligned}$$

Thus a reaction of price of one year on the supply situation of the next does not tend to produce any difference between long time and short time elasticity of demand. It does make a difference, however, in long and short time elasticities of supply. In the case of potatoes substitution of values already found gives  $e_L = +.52$  as the elasticity of the long time supply curve, insofar as determined by the reaction of the price of one year on the supply of the next. If it were legitimate to assume that there is no elasticity of supply within a year ( $e=0$ ), the formula for  $e_L$  reduces to  $S, \frac{\sigma_Q}{\sigma_P} = r_{QP'} \frac{\sigma_Q}{\sigma_P} = e_{QP'}$ .

### *Tests of Significance*

In considering the reliability of path coefficients there are two questions which must be kept distinct. First is the adequacy of the qualitative scheme to which the path coefficients apply and second is the reliability of the coefficients, if one accepts the scheme as representing a valid point of view. The setting up of a qualitative scheme depends primarily on information outside of the numerical data and the judgment as to its validity must rest primarily on this outside information. One may determine from standard errors whether the observed correlations are compatible with the scheme and thus whether it is a possible one, but not whether it correctly represents the causal relation.

Having accepted a certain scheme with which the data are compatible, one would like to determine the reliability of the values reached for the path coefficients. Obviously no single formula can be given, applicable to all cases. The basic formulae of the method are ones for writing series of simultaneous equations, which must be solved to obtain the unknown path coefficients and correlation coefficients. These equations are in general non-linear with respect to the unknown quantities, making it impossible to express the solution in a general formula in which substitution can be made in routine fashion.

Certain principles can, however, be illustrated by the results

in simple cases. No attempt will be made here to deal with the complications due to small numbers. It will be assumed that the errors of sampling are in general so small in comparison with the values of the coefficients that second degree terms in the errors may be ignored. It is recognized that a more thorough treatment of the matter is much to be desired.

The simplest set up (Fig. 33) is that in which one variable  $V_o$  is represented as a function of another  $V_i$ , and of a residual factor  $V_u$ . The equations are as follows:

$$(74) \quad P_{oi} = r_{oi}$$

$$(75) \quad P_{oi}^2 + P_{ou}^2 = 1 \quad . \text{ From (74) ,}$$

$$(76) \quad \sigma_{P_{oi}}^2 = \sigma_{r_{oi}}^2 = \frac{1 - r_{oi}^2}{N} .$$

From (75)

$$(77) \quad 2 P_{oi} \delta P_{oi} + 2 P_{ou} \delta P_{ou} = 0 ,$$

(assuming as noted above that  $\delta P_{oi}$  and  $\delta P_{ou}$  are small compared with  $P_{oi}$  and  $P_{ou}$  ).

$$(78) \quad \delta P_{ou} = - \frac{P_{oi}}{P_{ou}} \delta P_{oi}$$

$$(79) \quad \sigma_{P_{ou}}^2 = \frac{P_{oi}^2}{P_{ou}^2} \sigma_{P_{oi}}^2 = \frac{r_{oi}^2 (1 - r_{oi}^2)}{N} .$$

The standard error of the residual path coefficient in a system in which one variable is represented as determined by a number of others (Fig. 34) may be derived similarly

$$(80) \quad \sigma_{P_{ou}}^2 = \frac{1}{N} r_{o(12 \dots m)}^2 [1 - r_{o(12 \dots m)}^2] .$$

Consider next the case in which variable  $V_o$  is a function of two uncorrelated variables  $V_i$  and  $V_z$ , and of residual factor  $V_u$ .

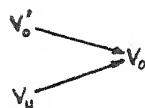


FIG. 33

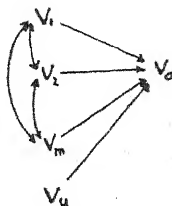


FIG. 34

Two different solutions are obtained for  $\sigma_{P_{01}}^2$  depending on the point of view. If it is accepted that  $V_1$  and  $V_2$  are wholly independent, except for the accidents of sampling, we have

$$(81) \quad P_{01} = r_{01}$$

$$(82) \quad \sigma_{P_{01}}^2 = \sigma_{r_{01}}^2 = \frac{(1 - r_{01}^2)^2}{N}$$

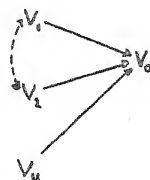


FIG. 35

If, however, there are no grounds for treating  $V_1$  and  $V_2$  as independent, except that  $r_{12}$  was insignificantly small in the data at hand, the proper set up is one in which a correlation between  $V_1$  and  $V_2$  is indicated as in Fig. 35.

$$(83) \quad r_{01} = P_{01} + P_{02} r_{12}$$

$$(84) \quad r_{02} = P_{01} r_{12} + P_{02}$$

giving

$$(85) \quad P_{01} = \frac{r_{01} - r_{02} r_{12}}{1 - r_{12}^2}$$

Treating sampling errors as differentials

$$(86) \quad \delta P_{01} = \frac{(1 - r_{12}^2)(\delta r_{01} - r_{02} \delta r_{12} - r_{12} \delta r_{02}) + 2(r_{01} - r_{02} r_{12}) r_{12} \delta r_{12}}{(1 - r_{12}^2)^2}$$

In the present case,  $r_{12}$  (but not  $\delta r_{12}$ ) is assumed to be zero in the sample at hand. Thus

$$(87) \quad \delta P_{01} = \delta r_{01} - r_{02} \delta r_{12}$$

$$(88) \quad \sigma_{P_{01}}^2 = \sigma_{r_{01}}^2 + r_{02}^2 \sigma_{r_{12}}^2 - 2 r_{02} m_{r_{01}, r_{12}},$$

where  $m_{r_{01}, r_{02}}$  is the product moment of deviations of  $r_{01}$  and  $r_{12}$ ,

$$(89) \quad m_{r_{01}, r_{02}} = r_{02} (1 - r_{01}^2)(1 - r_{12}^2) - \frac{r_{01} r_{12}}{2} (1 - r_{01}^2 - r_{02}^2 - r_{12}^2 + 2 r_{01} r_{02} r_{12})$$

by the formula of Pearson and Filon. Again treating  $r_{12}$  as negligibly small,

$$(90) \quad m_{r_{01}, r_{02}} = r_{02} (1 - r_{01}^2)$$

$$(91) \quad \sigma_{P_{01}}^2 = \frac{1}{N} [(1 - r_{01}^2)^2 + r_{02}^2 - 2 r_{02} (1 - r_{01}^2)]$$

$$(92) \quad = \sigma_{r_{01}}^2 - \frac{r_{02}^2 (1 - 2 r_{01}^2)}{N}$$

This is smaller than the value of  $\sigma_{P_{01}}^2$  obtained on the assumption of independence of  $V_1$  and  $V_2$  if  $r_{01}^2$  is less than  $1/2$ , but larger for larger values of  $r_{01}^2$ .

If the correlation between the two known factors  $V_1$  and  $V_2$  of figure 35, is not negligible, the squaring of the full formula for  $\sigma_{P_{01}}^2$  and division by  $N$ , leads after some reduction to the formula

$$(93) \quad \sigma_{P_{01}}^2 = \frac{1}{N} \left[ \frac{1 - r_{0(12)}^2}{1 - r_{12}^2} - P_{01}^2 (1 + r_{01}^2 - 2 r_{0(12)}^2) \right].$$

A somewhat rough estimate of the standard errors in the analysis of birth weight of guinea pigs, *page 179*, can be made by this formula. The correlation between birth weight and size of litter was however based on larger numbers (3353) than the correlation involving gestation period (1317). Adopting the smaller numbers we find

$$P_{BL} = -.51 \pm .020$$

$$P_{BG} = +.30 \pm .022.$$

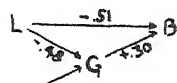


FIG. 36

While these estimates of the standard errors do not take cognizance of the approximation involved in substitution of gestation period for observed interval between litter (estimated  $r_{GL} = .95$ ) they are sufficient to indicate that the calculated path coefficient can be relied upon as accurate to a first order, assuming the correctness of the set up.

The standard error of a path coefficient has not been worked out for systems in which one variable is represented as affected by more than two known variables. The standard error of the closely allied concrete regression coefficient is however well known and can be used in testing significance.

$$(94) \quad \sigma_{c_{01}}^2 = \frac{\sigma_{0.12 \dots m}^2}{\sigma_{1.2 \dots m}^2} = \frac{\sigma_0^2 [1 - r_{0(12 \dots m)}^2]}{\sigma_1^2 [1 - r_{1(2 \dots m)}^2]}.$$

Since  $P_{01}^2 = c_{01}^2 \frac{\sigma_1^2}{\sigma_0^2}$ , the variance of the path coefficient can be written



$$(95) \quad \sigma_{P_{o1}}^2 = \frac{1 - r_{o(12 \dots m)}^2}{1 - r_{1(2 \dots m)}^2}, \text{ if } \frac{\sigma_o^2}{\sigma_1^2} \text{ can be treated}$$

as constant. This probably gives fairly good approximation in any case and is so used by Brandt (1928). In the case of guinea pig weight discussed above, the correct formula gives a result a little smaller than this approximation.

It will be noted that the standard errors may take very high values if the independent variable under consideration ( $V_i$ ) approaches complete determination by the others in the system, i.e. if  $[1 - r_{i(2 \dots m)}^2]$  approaches 0. In general, coefficients for paths leading from variables closely correlated with each other are subject to large standard errors. In making up a system, whether for prediction purposes or interpretation the aim should be to select factors closely correlated with the dependent variable but as nearly independent of each other as practicable.

If the dependent variable is completely determined by the specified factors ( $r_{o(12 \dots m)}^2 = 1$ ) the standard error of the concrete partial regression coefficient becomes zero. This is not the case with that of the path coefficient. Thus in the two factor case discussed above

$$(96) \quad \sigma_{P_{o1}}^2 = \frac{P_{o1}^2 (1 - r_{o1}^2)}{N} = \frac{(1 - r_{o1}^2)(1 - r_{o2}^2)}{N(1 - r_{12}^2)} \text{ if } r_{o(12)}^2 = 1.$$

More generally, if  $C_{o1}$  can be treated as constant (as it can if  $r_{o(12 \dots m)}^2 = 1$ ),

$$(97) \quad \sigma_{P_{o1}}^2 = C_{o1}^2 \sigma_{\left(\frac{C_1}{C_o}\right)}^2 = C_{o1}^2 \frac{\sigma_1^2}{\sigma_o^2} \frac{(1 - r_{o1}^2)}{N} = \frac{P_{o1}^2 (1 - r_{o1}^2)}{N}$$

which is in agreement with the preceding result.

Another simple set up, which is of interest is that in which three variables are arranged in chain sequence ( $r_{o2}^2 = r_{o1}^2 \cdot r_{12}^2$ ). Here again the point of view makes a difference. If the above relation is merely an empirical one, the situation is mere-

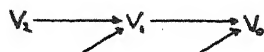


FIG. 37

ly a special case of that just discussed (the case in which  $P_{02} = 0$ , and  $P_{01} = r_{01}$ ). By substitution we find

$$(98) \quad \sigma_{P_{01}}^2 = \frac{1}{N} \left[ \frac{1-r_{01}^2}{1-r_{12}^2} - r_{01}^2 (1-r_{01}^2) \right]$$

$$(99) \quad \sigma_{P_{12}}^2 = \frac{1}{N} [1-r_{12}^2]^2$$

$$(100) \quad \sigma_{P_{02}}^2 = \frac{1}{N} \left[ \frac{1-r_{01}^2}{1-r_{12}^2} \right].$$

If, however,  $V_1$  is represented as the sole intermediary between  $V_0$  and  $V_2$  on theoretical grounds, the result is different. Two different determinations can be made of  $\sigma_{P_{01}}^2$  and of  $\sigma_{P_{12}}^2$ , the reason being that more equations can be written than there are unknown path coefficients. From,  $P_{01} = r_{01}$ ,

$$(101) \quad \sigma_{P_{01}}^2 = \frac{1}{N} (1-r_{01}^2)^2 \quad . \text{ From } P_{01} = \frac{r_{02}}{r_{12}},$$

$$(102) \quad \sigma_{P_{01}}^2 = \frac{1}{N} \frac{(1-r_{01}^2)(1-r_{02}^2)}{r_{12}^2} \quad . \text{ From } P_{12} = r_{12},$$

$$(103) \quad \sigma_{P_{12}}^2 = \frac{1}{N} (1-r_{12}^2)^2 \quad . \text{ From } P_{12} = \frac{r_{02}}{r_{01}},$$

$$(104) \quad \sigma_{P_{12}}^2 = \frac{1}{N} \frac{(1-r_{12}^2)(1-r_{02}^2)}{r_{01}^2}.$$

Similarly two determinations can be made of  $\sigma_{P_{012}}^2$ . From  $P_{012} = r_{02}$ ,

$$(105) \quad \sigma_{P_{012}}^2 = \frac{1}{N} [1-r_{02}^2]^2 \quad . \text{ From } P_{012} = r_{01} r_{12},$$

$$(106) \quad \sigma_{P_{012}}^2 = \frac{1}{N} [(1-r_{02}^2)^2 - (1-r_{01}^2)(1-r_{12}^2)].$$

With standard deviations calculated from two independent sets of data in each case, a combination estimate, smaller than either can be obtained from the formula

$$(107) \quad \sigma_{\text{Total}}^2 = \frac{1}{\sum \frac{1}{\sigma^2}}.$$

This illustrates the important principle that where there is a superfluity of equations for determining the path coefficients, the standard errors of these are correspondingly reduced. In the analysis of corn and hog correlations 42 path coefficients were found with which 510 correlations (and 4 cases of complete determination) were in agreement to the extent expected from their standard errors. Calculation of the standard errors of the path coefficients in this system seems out of the question, but it may safely be assumed that values of the order  $\frac{1}{\sqrt{N}}$ , which might be based on 42 equations are to be reduced by considerable amounts by the superfluity of data available.

There are some interesting contrasts in the standard errors given above. If  $\chi_{12}$  is large,  $\sigma_{P_{01}}^2$  may be large in the empirical system. But if the theory that  $\chi_1$  is the only intermediary rests on adequate grounds, independent of the observed correlations,  $\sigma_{P_{01}}^2$  may be small with large  $\chi_{12}$ .

We will conclude with consideration of a set up like that used for the relation of supply and demand to price and quantity. It will be assumed first that the number of cases is large (a condition contrary to that found in the examples given). Differentiation of the 5 basic equations gives 5 equations expressing the relations between small deviations of the path coefficients and correlations.

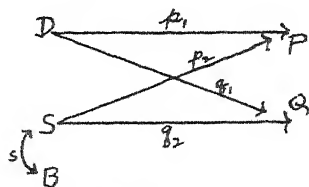


FIG. 38

(108) $p_1^2 + p_2^2 = 1$	(113) $2 p_1 \delta p_1 + 2 p_2 \delta p_2 = 0$
(109) $q_1^2 + q_2^2 = 1$	(114) $2 q_1 \delta q_1 + 2 q_2 \delta q_2 = 0$
(110) $p_1 q_1 + p_2 q_2 = \chi_{PQ}$	(115) $p_1 \delta q_1 + q_1 \delta p_1 + p_2 \delta q_2 + q_2 \delta p_2 = \delta \chi_{PQ}$
(111) $p_2 s = \chi_{BP}$	(116) $p_2 \delta s + s \delta p_2 = \delta \chi_{BP}$
(112) $q_2 s = \chi_{BQ}$	(117) $q_2 \delta s + s \delta q_2 = \delta \chi_{BQ}$

Thus

$$(118) \quad \delta p_1 = -\frac{p_2}{p_1} \delta p_2$$

$$(119) \quad \delta q_1 = -\frac{q_2}{q_1} \delta q_2$$

$$(120) \quad \delta q_2 (p_2 - \frac{q_2}{q_1} p_1) + \delta p_2 (q_2 - \frac{p_2}{p_1} q_1) = \delta r_{PQ}$$

$$(121) \quad \delta q_2 \cdot p_2 - \delta p_2 \cdot q_2 = \frac{p_2}{s} \delta r_{BQ} - \frac{q_2}{s} \delta r_{BP}$$

Solution of (120) and (121) as simultaneous equations gives expressions for  $\delta p_2$  and  $\delta q_2$  in terms of  $\delta r_{PQ}$ ,  $\delta r_{BQ}$  &  $\delta r_{BP}$ , from which their squared, standard errors can be found by taking the average squares. Letting  $A$ ,  $B$  and  $C$  be the coefficients,

$$(122) \quad \delta q_2 = A \delta r_{PQ} + B \delta r_{BQ} + C \delta r_{BP}$$

$$(123) \quad \sigma_{q_2}^2 = A^2 \sigma_{r_{PQ}}^2 + B^2 \sigma_{r_{BQ}}^2 + C^2 \sigma_{r_{BP}}^2 + 2AB m_{r_{PQ} r_{BQ}} + 2AC m_{r_{PQ} r_{BP}} + 2BC m_{r_{BQ} r_{BP}}$$

The product moments of the deviations of the correlation coefficients can be found by Pearson & Filon's formula cited on page 206.

The standard errors of  $\delta p_1$  and  $\delta q_2$  can be found at once with the help of equations (11) and (12) while that of  $\delta s$  can be found from (9) or (10) after expressing  $\delta p_2$  (or  $\delta q_2$ ) in terms of deviations of the known correlation coefficients.

The significance of the coefficients of elasticity is most easily investigated by taking these on scales in which the standard errors of the percentage deviation in price and quantity are taken as unity i.e. by finding the standard error of  $\frac{q_1}{p_1}$  and of  $\frac{q_2}{p_2}$  instead of  $e = \frac{q_1 \sigma_q}{p_1 \sigma_p}$  and  $\eta = \frac{q_2 \sigma_q}{p_2 \sigma_p}$  respectively. These standard errors can be found from the formula for the standard error of a ratio.

$$(124) \quad \sigma_{\frac{q_1}{p_1}}^2 = \frac{q_1^2}{p_1^2} \left[ \frac{\sigma_{q_1}^2}{q_1^2} + \frac{\sigma_{p_1}^2}{p_1^2} - 2 \frac{m_{q_1 p_1}}{q_1 p_1} \right]$$

The product moments of the path coefficients can be obtained

by squaring equation (8) after expressing  $\delta p_1$  and  $\delta g_1$  in terms of  $\delta p_2$  and  $\delta g_2$  equation (13) or the converse.

The numbers of cases in the actual examples were not large enough to make the method a satisfactory one. The calculations have been carried through, however, with the results given below.

	Summer Pork 26 Years	Winter Pork 44 Years	Potatoes 19 Years
$r_{PQ}$	$-.63 \pm .12$	$-.68 \pm .08$	$-.85 \pm .06$
$r_{PB}$	$-.47 \pm .16$	$-.63 \pm .09$	$-.56 \pm .16$
$r_{QB}$	$-.64 \pm .12$	$+.83 \pm .05$	$+.65 \pm .14$
$p_1$	$+.69 \pm .20$	$+.66 \pm .11$	$+.50 \pm .26$
$p_2$	$-.73 \pm .19$	$-.76 \pm .09$	$-.86 \pm .15$
$g_1$	$+.13 \pm .24$	$+.11 \pm .10$	$+.02 \pm .27$
$g_2$	$+.99 \pm .03$	$+.99 \pm .01$	$+1.00 \pm .00$
$S$	$+.65 \pm .12$	$+.84 \pm .05$	$+.65 \pm .14$
$g_1/p_1$	$+.19 \pm .39$	$+.16 \pm .17$	$+.05 \pm .57$
$g_2/p_2$	$-1.36 \pm .38$	$-1.32 \pm .17$	$-1.16 \pm .21$
$e$	$+.13$	$+.11$	$+.03$
$\eta$	$-.94$	$-.88$	$-.82$

The most nearly satisfactory case is that of winter pork based on rather large primary correlations obtained from 44 years' experience, but even here, the standard error of  $g_1$  is nearly as

large as  $g_i$  itself. In the other cases, the standard error of  $g_i$  is larger than  $g_i$ . The term  $\delta g_i^2$  omitted in equation (7) is of the order of the term  $x_{gi} \delta g_i$ , or larger, making this equation invalid. The other equations are not affected, at least to anything like as great an extent. An approximate solution can be obtained even though equation (7) is omitted, from the consideration that in this case in which  $g_i$  is small,  $\delta g_i$  must be very small and may be ignored. Thus

$$(125) \quad \delta s = \delta r_{BQ}$$

$$(126) \quad \delta p_2 = \frac{1}{s} \left[ \delta r_{GP} - \frac{r_2}{g_2} \delta r_{BQ} \right]$$

$$(127) \quad \delta g_1 = \frac{r_1}{f_1} \left[ \delta r_{PA} - g_2 \delta p_2 \right]$$

$$(128) \quad \delta p_1 = - \frac{r_2}{f_1} \delta p_2.$$

The results are substantially the same as those obtained above, since the values assigned  $\delta g_2$  were very small, even if not reliable. It may be safely concluded that winter pork at the large markets has very little elasticity of supply but a moderate elasticity of demand. The results for summer pork and for potatoes are in harmony with similar interpretations but are based on such inadequate numbers as to have little significance in themselves.

#### REFERENCES

- Brandt, A. E., 1928—Calculation and use of the standard deviation of partial regression coefficients. *Iowa St. College Jour. Sci.* 2: 235-242.
- Burks, B. S., 1928—The relative influence of nature and nurture upon mental development; a comparative study of foster-parent—foster-child resemblance and true parent—true child resemblance. 27th Yearbook of Nat. Soc. for Study of Education, 1928, Part I: 219-316.
- Calder, A., 1927—The role of inbreeding in the development of the Clydesdale breed of horse. *Proc. Roy. Soc. Edinb.* 47: 118-140.
- Fisher, R. A., 1925—Statistical methods for Research Workers. 239 pp. Oliver & Boyd. Edinburgh.
- Jennings, H. S., 1916—The numerical results of diverse systems of breeding. *Genetics* 1: 53-89.
- Kelley, F. L., 1927—Statistical Method. 390 pp. The Macmillan Co. New York.
- Krichewsky, S., 1927—Interpretation of Correlation Coefficients. Ministry of Public Works. Egypt. Physical Dept. Paper No. 22, Cairo.

- Lush, J. L., 1930—The number of daughters necessary to prove a sire. *Jour. Dairy Sci.* 13: 209-220.
- 1932—The amount and kind of inbreeding which has occurred in the development of breeds of livestock. *Proc. 6th Internat. Congress of Genetics* 2: 123-126.
- McPhee, H. C., and S. Wright, 1925—Mendelian analysis of the pure breeds of live stock. III. The Shorthorns. *Jour. Hered.* 16: 205-215.
- 1926—Mendelian analysis of the pure breeds of live stock. IV. The British Dairy Shorthorns. *Jour. Hered.* 17: 397-401.
- Minot, C. S., 1891—Senescence and rejuvenation. *Jour. Physiol.* 12: 97-153.
- Niles, H. E., 1922—Correlation, Causation and Wright's theory of path coefficients. *Genetics* 7: 258-273.
- 1923—The method of path coefficients, an answer to Wright. *Genetics* 8: 256-260.
- Smith, A. D. B., 1926—Inbreeding in cattle and horses. *Eugen. Rev.* 14: 189-204.
- Wright, P. G., 1928—The tariff on animal and vegetable oils. 347 pp. The Macmillan Co., New York.
- Wright, S., 1918—On the nature of size factors. *Genetics* 3: 367-374.
- 1920—The relative importance of heredity and environment in determining the piebald pattern of guinea pigs. *Proc. Nat. Acad. Sci.* 6: 320-332.
- 1921a—Correlation and Causation. *Jour. Ag. Res.* 20: 557-585.
- 1921b—Systems of mating. *Genetics* 6: 111-178.
- 1922—Coefficients of inbreeding and relationship. *Am. Nat.* 56: 330-338.
- 1923a—The theory of path coefficients—a reply to Niles' criticism. *Genetics* 8: 239-255.
- 1923b—Mendelian analysis of the pure breeds of livestock. I. The measurement of inbreeding and relationship. *Jour. Hered.* 14: 339-348. II. The Duchess family of Shorthorns as bred by Thomas Bates. *Jour. Hered.* 14: 405-422.
- 1925a—Corn and hog correlations. *Bull. No. 1300*, 60 pp. U. S. Dept. of Agric.
- 1926a—A frequency curve adapted to variation in percentage occurrence. *Jour. Amer. Stat. Assoc.* 21: 162-178.
- 1926b—Effects of age of parents on characteristics of the guinea pig. *Amer. Nat.* 60: 552-559.
- 1927—The effects in combination of the major color-factors of the guinea pig. *Genetics* 12: 530-569.
- 1931a—Statistical methods in biology. *Jour. Amer. Stat. Ass. Supplement. Papers and Proceedings of the 92nd annual meeting.* 26: 155-163.
- 1931b—Evolution in mendelian populations. *Genetics* 16: 97-159.
- 1932a—On the evaluation of dairy sires. *Proc. Amer. Soc. Animal Prod.* 1932: 71-78.

- 1932b—General, group and special size factors. *Genetics* 17: 603-619.
- Wright, S., and H. C. McPhee, 1925—An approximate method of calculating coefficients of inbreeding and relationship from livestock pedigrees. *Jour. Ag. Res.* 31: 377-383.
- Wright, Sewall, 1933a—Inbreeding and homozygosis. *Proc. Nat. Acad. Sci.* 19: 411-420.
- 1933b—Inbreeding and recombination. *Proc. Nat. Acad. Sci.* 19:420-433.

*Sewall Wright*



# MATHEMATICAL FOUNDATION FOR A METHOD OF STATISTICAL ANALYSIS OF HOUSEHOLD BUDGETS

By JOHN W. BOLDYREFF  
*Harvard University*

The object of this paper is to offer a satisfactory method of statistical analysis of household budgets in accordance with the general principles of mathematical logic. I have, therefore, taken these words of Fourier: "Mathematics has no symbols for confused ideas"<sup>1</sup> as my guiding light, and set out to effect a simple and comprehensive analysis of the general type of statistical data which is included under the heading "household budgets," i. e. monetary incomes and expenditures of these incomes.

I have tried to lay the greatest stress, accordingly, on the clarity and terseness of the exposition rather than inclusiveness, attempting to diminish to the utmost the number of undefined ideas and the undemonstrated propositions. I make no special claim to originality and base my method upon the works of numerous previous investigators, summarizing analytically old principles and ideas on the bases of mutual consistency and reducibility to more fundamental principles. This paper is specially framed to relieve the feeling of intellectual discomfort which of late has been troublesome to conscientious investigators in our field, so overcrowded with revelations of numerous parts, rather than with indications of the mode of combination of the major components within the whole. I address here the properly instructed mind and so dispense at times with the elaboration of some statements.

In this summary, therefore, we shall be concerned with laying down a rigid method for analysing the budgetary data, defining their scope formally to include only the monetary incomes and

---

<sup>1</sup> Quoted from J. A. Schumpeter, *Die Wirtschaftstheorie der Gegenwart*, Wien, I, 11, 1927.

the relative amounts of these incomes spent in a defined manner. This, naturally, excludes all reference to economic theory (e. g. utility, demand curves, etc.) from our discussion; I do so not because of a desire to depreciate the importance of that kind of belief, but because I do not wish to consider it here.

Obviously, "it is never a mathematical proposition which we need, but we use mathematical propositions *only* in order to infer from propositions which do not belong to mathematics to others which equally do not belong to mathematics."<sup>2</sup> Moreover, it is also true that nothing can be purely logical or mathematical (unless we follow Hilbert and define mathematics as a game with meaningless marks on paper); all propositions involve some psychological terms such as defining, meaning, asserting or naming. The method and scope of a mathematical analysis is in a like manner dependent on the purpose for which it is to be undertaken.

The purposes in the study of budgetary data assume varying emphasis depending on the point of view of approach, that of economics, home economics, social welfare, and sociology.<sup>3</sup> All of these approaches are concerned with the relation between the sizes of incomes and the relative amounts spent for certain goods and services.

Generally, the classification of expenditures of an income is made as to the amounts (or proportions) spent for food, clothing, rent, light, education, health, recreation, savings, and amusement. Some investigators limit their classifications to five items: food, clothing, rent, fuel and light, and sundries (everything not included under the first four). Others prefer to subdivide the classification further and break up each of the above nine types of expenditure into what they deem to be its component parts, and proceed to study these new relationships and to generalize from them. On my part, I judge the latter performances ex-

---

<sup>2</sup> Wittgenstein, *Tractatus Logico-Philosophicus*, 6, 211.

<sup>3</sup> C. C. Zimmerman, *Am. J. Soc.*, vol. XXXIII, 6, 1928.

tremely dangerous. It seems to me, that the analysis of the major components of income's expenditure in their relationships to the size of income and to each other should be developed and perfected beforehand, and then only gradually extended to apply to the minor items. Moreover, the splitting up of a few variables (the types of expenditures) into many introduces other difficulties—aside from the fact that a study of simple relationships is apt to be more clarifying—the introduction of a component part of the whole variable as a new variable, immediately raises the question why this component is isolated and not the other. None of the arguments that can be generally cited (and usually no arguments are cited) are really decisive, and the position is extremely unsatisfactory to anyone with real curiosity about the fundamental relationships. Unless we wish the analysis of the budgetary data to remain self-contradictory and meaningless, we must adopt a limiting method, and study not more than two variables at a time. Then, and only then, can we hope to establish or discover any "laws," or functional relationships.

In my experiments to develop a satisfactory method of analysis I would begin generally with five classes of expenditures: food, clothing, rent, fuel and light, and sundries. Later, I have come to the conclusion that some of these tend to have a sort of complementary relationship between them. Thus, "fuel and light" are often higher or lower with a higher or lower "rent," and in some cases a part of "rent" covers "fuel and light," in other cases the discomfort and monetary cost of "fuel and light" lowers the "rent" expenditure. Likewise, some complementary relationship is observed between "fuel and light" and "clothing" (especially in submarginal households) and between "clothing" and "rent" (e. g. social demand of the stylish residential district). These are merely a few examples which led me to question the validity of initial isolation of these three items (clothing, fuel and light, and rent) from each other. Accordingly, I suggest to limit our

investigation to the study of possible relationships between: (1) the size of the income and (a) the amount spent for food, (b) the amount spent for sundries; (2) the amount spent for food and the amount spent for sundries—assuming temporarily for convenience and analysis all other itemized expenditures under rent, clothing, and fuel and light, to be not subject to individual isolation.

As to the unit in household budget, the variety of units employed bewilders at first a mathematical student. Of these, the old scale of two children for one adult, the various other "adult equivalents" (e.g. Engel's quet scale of 3.0 for woman of 20 and 3.5 for man of 25 years; Atwater's scale of 10, 8, 7, 5, 2.5; then the scales of Voit, U.S.D. of L., H. C. Sherman and L. H. Gillett, G. Lusk, L. Emmett Holt, and others—each scale giving "adult equivalents" for children, male and female), all clearly show inability of investigators to agree on a scale to determine the size of a family in standard units. It seems to me that the inventors of such scales forget somehow that "taking an arbitrary individual in the living nature—a man, an animal, a plant—it will generally be found impossible to find out another individual in all respects identical to the first one chosen."<sup>4</sup> The standard scale in budgetary studies is less valid than usual statistical abstractions, for such factors as geographic space (climate, nutritive ratio, energy value, cost), social space (stratification and differentiation), economic space (size of incomes), occupational space (caloric requirement, etc.), time factor (daily, weekly, monthly, seasonal, and longer fluctuations), as well as age (for there is a great latitude in "adult" ages and a corresponding variability in "requirements") and sex differences, are admittedly affecting each budgetary individual in a variety of unknown ways. In view of the complexity of the problem and the enormousness of human population, any "adult equivalent" scale will appear

---

<sup>4</sup>C. V. L. Charlier, *Acta Universitatis Ludensis*, 1905-6, XVI, 5, p. 3.

to be based on samples obtained in gross violation of the sampling theory, for it is very doubtful that a sufficiently large and representative sample can be secured and it is very hard to see how it can escape being greatly biased. Besides, most of these scales are based on energy requirement only, and refer to "food" but not at all to other types of expenditure; therefore, they would be of little general significance even if they were valid in their specific aspect. Personally, I must reject all such scales as meaningless and incline to hesitate between adopting a "normal family" (on basis of a standard number of members, irrespective of their characteristics) and a "household" (irrespective of number of members and of their characteristics), the presumption being that in a sufficient random sample the differences either way will tend to cancel out. This may not seem to be a more accurate method than others, but, in all probability, it is just as accurate, and its virtue lies, moreover, in the fact that its limitations are all on the surface instead of being hidden away behind a misleading label. The data of the last Census seem to favor this attitude.<sup>5</sup>

The purpose of budgetary analysis is to discover, allegedly, certain functional relationships, if any, between the varying income and the relative amounts of each type of expenditure. To discover such relationships and to determine them explicitly one must recognize that all laws logically function within limits. One needs not go as far as Hilbert and insist that anything involving an infinity of any kind must be meaningless—in pure mathematics this may be a useful abstraction—but it should be obvious that in all organic laws anything infinite appears a stupid fiction which cannot be argued for except by proceeding to a limit. The behavior of the budgetary items is clearly a biotic phenomenon which fact some of the investigators in our field tend to overlook consistently. If there are any functional relationships in the bud-

---

<sup>5</sup> L. E. Truesdell, *New Family Statistics for 1930*, J. Am. Statist. Assn., March 1933 (Supplement), pp. 154-8.

getary data these will be found only within definite limits of minimum and maximum, and any contradicting evidence to such laws if found below or above these limits cannot be interpreted as disproving such laws.

We shall make our points clearer by illustrating the above exposition by the so-called Engel's Law (I am referring to the second part of it), incidentally commenting briefly on its validity and demonstrating the details of our method.

It will not be amiss to formulate in a few words the part of Engel's Law (1895) we shall be concerned with in our discussion. Comparing the incomes of laboring families, middle class families, and well-to-do families, Engel conjectured that:

- (1) the greater the income, the smaller the percentage of outlay for subsistence (food),
- (2) percentage of outlay for clothing is approximately the same, whatever the income,
- (3) percentage of outlay for rent, and for fuel and light, is approximately the same, whatever the income,
- (4) as income increases in amount, the percentage of outlay for sundries becomes greater.

Most of the investigators incline to accept the first and the last of Engel's propositions, both from the static and dynamic viewpoints. As for myself, I like to consider this law with reference to the following questions:

(1) as incomes increase does the percentage of outlay for food decline and the percentage of outlay for sundries increase?

(2) is this a static law; i.e. in a given place, at a given time, will there be a higher percentage of outlay for sundries and lower percentage of outlay for food with larger incomes, and vice versa for smaller incomes?

(3) does this hold in the dynamic aspect—as incomes increase (in time) do the percentages of outlay for food

decline and those for sundries rise, for short and long time?

(4) is this law reversible, i. e. if incomes decrease do the percentages of outlay for food rise and those for sundries decline, statically and dynamically?

(5) can the percentages of outlay for clothing, rent, and fuel and light be treated as constant, statically and dynamically?

(6) can this law be interpreted to mean that when the percentage of outlay for food declines the percentage of outlay for sundries rises, and vice versa, statically and dynamically?

(7) if this law is valid, what is its significance for forecasting?

Let us consider first the problem of limits from a purely abstract viewpoint. We assume for the sake of argument this law to be valid and set up a hypothetical series of incomes with the respective percentages and amounts of outlays for food and for sundries. The following example shows clearly that a limit is eventually reached when the law becomes automatically in-operative.

TABLE I.

<i>Income in \$</i>	<i>% for Food</i>	<i>\$ for Food</i>	<i>% for Sundries</i>	<i>\$ for Sundries</i>
Under 900	—	—	—	—
" 1,000	50	500	10	100
" 2,000	45	900	15	300
" 3,000	40	1,200	20	600
" 4,000	35	1,400	25	1,000
" 5,000	30	1,500	30	1,500
" 6,000	25	1,500	35	2,100
" 7,000	20	1,400	40	2,800
" 8,000	15	1,200	45	3,600
" 9,000	10	900	50	4,500
" 10,000	5	500	55	5,500

Aside from demonstrating the inevitableness of limits, this illustration shows also that from purely common sense considerations constancy of interrelationship between variation of percentages for food and percentages for sundries is not feasible. That the absolute amount spent for food cannot decline with increase of income but should constantly keep on rising (though, perhaps, in small amounts), should be clear from common sense, even if we shall consider this amount as stationary after a certain sum is reached and credit the increase to sundries (cooks, maids, travel, eating out, etc.)—yet, even in such cases decline should be out of the question.

Now we can give an illustration of the validity of assumption that the percentages of outlay for rent, clothing, and fuel and light, for convenience of analysis and until proven to be contrary, can be held constant. We have tried this with a variety of data and generally found this to be true.

TABLE II.

COMPARISON OF THE PERCENTAGES OF THE TOTAL FAMILY EXPENDITURE FOR THE DIFFERENT GROUPS OF LIVING COSTS<sup>6</sup>

<i>Item</i>	<i>Eden's 73 English Budgets 1796</i>	<i>Engel's Belgian Data 1853</i>	<i>Le Play's Method 100 Budgets 1829-88</i>	<i>U.S.D.L. (2562) 1890-1</i>	<i>U.S.D.L. (12096) 1918-9</i>	<i>Groton, N. Y. (92) 1919</i>
Food	73	66.9	56.8	41.1	38.2	41.7
Rent	12	7.6	6.8	15.1	13.4	13.1
Clothing	7	14.9	16.5	15.3	16.6	11.3
Fuel and light	5	5.6	4.3	5.9	5.3	6.8
Sundries	3	5.0	15.6	27.7	26.4	27.1

Adding the "rent" and "clothing" items from Table II we obtain: 19.0, 22.5, 23.3, 30.4, 30.0, and 24.4; by adding to these their respective "fuel and light" items we obtain: 24.0, 28.1,

<sup>6</sup> Taken from *Noble*, Cornell University Agricultural Experiment Station Bulletin, # 431, Sept., 1924.



27.6, 36.3, 35.3, and 31.2. It seems justifiable to assume these items in their summation to be a constant factor in time analysis. That they are constant for static analysis will be shown later. But it may be mentioned in passing that taking the data from Noble's Table 19 (average percentages of expenditure of items of cost of living of 518 families in New York City, by income groups)<sup>7</sup> and adding up our "constant factor" we get 35.9 for the lowest income group and 36.4 for the highest.

One illustration more. Below are the figures taken from the U. S. B. L., 18th annual report, 1904, p. 101.

TABLE III.

<i>Classified Income</i>	<i>Rent</i>	<i>Fuel</i>	<i>Light</i>	<i>Food</i>	<i>Clothing</i>	<i>Sundries</i>
Under \$ 200	16.93	6.69	1.27	50.85	8.68	15.58
" 300	18.02	6.09	1.13	47.33	8.66	18.77
" 400	18.61	5.97	1.14	48.09	10.02	16.09
" 500	18.57	5.54	1.12	46.88	11.39	16.50
" 600	18.43	5.09	1.12	46.16	11.98	17.20
" 700	18.48	4.65	1.12	43.48	12.88	19.39
" 800	18.17	4.14	1.12	41.44	13.50	21.63
" 900	17.07	3.87	1.10	41.37	13.57	23.02
" 1,000	17.58	3.85	1.11	39.90	14.35	23.21
" 1,100	17.53	3.77	1.16	38.79	15.06	23.69
" 1,200	16.59	3.63	1.08	37.68	14.89	26.13
1,200 and over	17.40	3.85	1.18	36.45	15.72	25.40

The "constant factor" taken at the lowest and highest incomes is found to be 33.57 and 36.15 respectively. The examination of the table from the point of view of finding a law, or functional relationship, reveals such phenomenon for the range of incomes from \$500 to \$1,200, inclusive. We shall proceed to examine the data included in these limits in accordance with our method.

We find the "constant factor" for \$500 income to be 36.62 and for \$1,200 income, 36.19.

<sup>7</sup> Op. cit.

We assumed a straight line relationship and computed simple coefficients of correlation between:

- (1) incomes and percentages of outlay for food
- (2) incomes and percentages of outlay for sundries
- (3) percentages of outlay for food and those for sundries

We want to stress in this connection that to us the coefficient of correlation means a measure of relationship which is already empirically established, not a proof of such relationship. We used L. P. Ayres<sup>8</sup> formula which we found convenient for computing purposes. To avoid a fictitious correlation between incomes and the percentages of outlay for food and for sundries, we have divided the income column by a constant. To facilitate computation we have likewise divided the "percentages for food" and the "percentages for sundries" columns by constants.

In making a summary comment on Engel's law, I would like to stress the following points from a purely methodological viewpoint. There seems to be definite evidence that in a given place, at a given time, the law holds consistently within certain limits. For very low income groups some other law may hold, or no law at all, and as to how extremely large incomes are spent we do not know. From the dynamic aspect, the law appears to have been working from the time of the French revolution up to the beginning of the present depression (much evidence could be cited to support this fairly well known fact, e.g. works of Schmoller, Rogers, D'Avenel, U. S. B. L. S. Bulletins, etc.). However, the study of W. A. Berridge (The need for a new survey of family budgets and buying habits, N. Y. Times, May 10, 1931, and "The Annalist," July 17, 1931) seems to indicate that from the secular standpoint this law is not immediately reversible, for with the shrinking incomes we observe a definite decline in the outlays

---

<sup>8</sup> *J. Educ. Research*, I, March-June, 1920.

for all items, including food, except for the outlay for sundries which appears to be almost stationary.

That percentages of outlay for clothing, rent, and fuel and light, can be added up and treated as a constant factor both statically and dynamically with rising incomes we can be reasonably certain of; what will happen with decreasing incomes in time analysis we are not ready to say. However, it must be borne in mind that even with the rising incomes the relationship between the percentages of outlay for food and for sundries need not be perfect as one may be led to think from their high individual coefficients of correlation with income in the example given above.

As to a practical application of the budgetary analysis to forecasting, I shall venture to say that in a socially planned society (if such society is workable), the study of itemized expenditures may prove invaluable. In other societies it may be used to forecast some sort of consumption indices—if these will be successfully computed they will undoubtedly help to flatten the curve of business cycles to an appreciable degree. As to how to develop these indices, I have no suggestion to make just now, except that it must be on basis of extension of a crude analysis similar to one offered here, and application of probability technique, properly based on psychological and historical findings. All I hope to have made clear in this paper is that the subject is very difficult, and that an analysis offered here is sufficient as a first step.

In conclusion, I must stress my indebtedness to Professors J. D. Black, J. A. Schumpeter, and C. C. Zimmerman for advice and suggestions. I am also grateful to Professor Zimmerman for the materials he let me examine. But above all I am indebted to Professor W. L. Crum from whom my point of view and method of attack are wholly derived; anything of value that I may have said in this paper is due to him.

John W. Boldynff

# ON THE RELATIVE STABILITY OF THE MEDIAN AND ARITHMETIC MEAN, WITH PARTICULAR REFERENCE TO CERTAIN FREQUENCY DISTRIBUTIONS WHICH CAN BE DISSECTED INTO NORMAL DISTRIBUTIONS<sup>1</sup>

By  
HARRY S. POLLARD

## I.

### THE CHOICE OF AN AVERAGE

In any statistical investigation in which an average is to be used as a summarizing figure for a frequency distribution the question arises, which average best describes the distribution. That this is still a debatable question among writers on economic statistics is shown by a perusal of the many papers dealing with the measurement of seasonal variations which have appeared in recent years.<sup>2</sup>

Each of the proposed methods of isolating seasonal variations involves an averaging, either of monthly items or of relatives of monthly items, but whether this averaging is best accomplished by use of the arithmetic mean, the median, or the mean of a middle group of items seems to be a moot point. Persons<sup>3</sup> employs the median of link relatives of monthly items, since by this device the influence of large non-seasonal variations may be greatly moderated. Hart<sup>4</sup> in justifying the use of the arithmetic mean has shown that the method of monthly means gives the actual

---

<sup>1</sup> A resume of a dissertation, bearing the same title, written under the direction of Professor Mark H. Ingraham and submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the University of Wisconsin, 1933.

<sup>2</sup> For a bibliography of literature on this subject see Mills, F. C., *Statistical Methods*, p. 343.

<sup>3</sup> Persons, W. M., *Correlation of Time Series*, Jour. Amer. Statis. Assn., June, 1923, p. 717.

<sup>4</sup> Hart, W. L., *The Method of Monthly Means for Determination of a Seasonal Variation*, Jour. Amer. Statis. Ass'n., Sept., 1922, pp. 341-349.

monthly values of the seasonal variation in case the seasonal variation is strictly periodic throughout the period of years under consideration and the long term variations are also periodic with integral numbers of years as their periods. The proof of this theorem is based on a property of Fourier series discussed by Bocher<sup>5</sup>.

The point of view of this paper is that another factor of importance should influence the choice of an average, that this choice should be guided not alone by consideration of exceptional cases which may arise, nor by theory which assumes a periodicity seldom found in sequences of economic data, but also by a consideration of the stability of the averages. For if a given frequency distribution is regarded as a random sample drawn from a theoretical distribution which contains a very large number of items, the accuracy with which a particular average of the sample will typify the entire theoretical distribution is influenced by the frequency curve for that average. It is the purpose of this paper to compare the stability of the arithmetic means and medians of frequency distributions which may be dissected into two and three normal distributions, and to develop a general method of comparing the relative stability of the mean and median which shall be applicable to any frequency distribution.

The dissection of a frequency curve into two normal components has been discussed by Karl Pearson<sup>6</sup>, who has developed methods for determining the values of the parameters of both symmetrical and asymmetrical frequency functions. He has applied these methods to distributions of cranial weights. Crum<sup>7</sup> has used Pearson's method of dissecting a symmetrical distribution in his discussion of the relative stability of the median and mean of link

---

<sup>5</sup> *Annals of Mathematics*, Second Series, vol. 7, p. 135, Formula (63).

<sup>6</sup> Pearson, K., *Contributions to the Mathematical Theory of Evolution*, *Philosophical Transactions*, Series A, vol. 185, 1894, pp. 71-110.

<sup>7</sup> Crum, W. L., *The Use of the Median in Determining Seasonal Variation*, *Jour. Amer. Statis. Ass'n.*, March, 1923, pp. 607-614.

relatives of monthly figures for the rate of interest on sixty to ninety-day commercial paper for the years 1890-1917, and his results are discussed in section VI of this paper. Our interest in an asymmetrical distribution composed of two normal distributions arises from the fact that such a distribution affords a good fit both to distributions which possess two distinct modes, and to skewed distributions with one mode. The study of a distribution which may be dissected into three normal components is suggested by the occurrence in economic data of tri-modal distributions. This paper will be concerned with only a particular class of three-component distributions, those which are symmetrical.

The hypothesis from which this investigation started was that a good criterion for measuring the stability of an average is its standard deviation. However, a difficulty which soon presented itself was the accurate determination of the standard deviation of the median. The classical formula for expressing the standard deviation,  $\sigma_M$ , of the medians of samples of  $s$  items each, drawn from a frequency distribution whose equation is  $y = f(x)$  and which satisfies the condition:

$$\int_{-\infty}^{\infty} f(x) dx = \frac{1}{k} = \int_0^{\infty} f(x) dx \quad \text{is} \quad \sigma_M = \frac{1}{2\sqrt{s} \cdot f(0)}.$$

The approximation to the value of the standard deviation of the median given by this formula is discussed in section IV, where it is shown that, although this approximation is close to the true value of the standard deviation of the median when  $s$  is large, it may be a very poor approximation when  $s$  is small, particularly for certain types of frequency curves.

Since it became obvious that the relative stability of the medians and arithmetic means of small samples cannot be determined by the methods which are valid when the samples are large, this paper resolved itself into two distinct investigations: a treatment of certain frequency functions using the classical formula for the standard deviation of the median, valid for large values of  $s$  ;

and the development of a second method of comparing the stability of the arithmetic mean and median which may be applied also when  $S$  is small. The first of these topics is considered in sections II and III, the second is taken up in sections IV and V, and is mathematically the more interesting part of the work. In section VI the various methods of comparing the stability of the arithmetic mean and median are applied to a particular sequence of economic data.

## II.

### THE RELATIVE MAGNITUDE AND STABILITY OF THE ARITHMETIC MEAN OF A FREQUENCY DISTRIBUTION WHICH IS COMPOSED OF TWO NORMAL DISTRIBUTIONS.

#### 1. *The Mean and Median and Their Standard Deviations.*

In this section a study will be made of the frequency function whose equation is

$$(1) \quad y = \frac{1}{\sqrt{2\pi}} \left( \frac{c_1}{\sigma_1} e^{-\frac{x^2}{2\sigma_1^2}} + \frac{c_2}{\sigma_2} e^{-\frac{(x-b)^2}{2\sigma_2^2}} \right),$$

with the purpose of determining the influence of the five parameters of this equation upon the location of the mean and median of the distribution, and upon the standard deviations of these averages.

The only conditions imposed upon the parameters  $c_1$ ,  $c_2$ ,  $\sigma_1$ ,  $\sigma_2$ ,  $b$  are that they shall assume only positive values (since they represent, respectively, the areas of the two component curves, their standard deviations, and the distance between their arithmetic means), and that the first two parameters shall satisfy the equation

$$(2) \quad c_1 + c_2 = 1,$$

so that the total probability, as represented by  $\int_{-\infty}^{\infty} y \, dx$ , shall be unity.

The arithmetic mean,  $\bar{x}$ , of the distribution may be expressed as a function of the parameters by the equation

$$(3) \quad \bar{x} = c_2 b.$$

The median,  $M$ , of the distribution satisfies the equation

$$(4) \quad \int_M^{\infty} \frac{1}{\sqrt{2\pi}} \left( \frac{c_1}{\sigma_1} e^{-\frac{x^2}{2\sigma_1^2}} + \frac{c_2}{\sigma_2} e^{-\frac{(x-b)^2}{2\sigma_2^2}} \right) dx = \frac{1}{2},$$

and can in general be located only by interpolation in a table of areas under the normal curve. This interpolation can be more easily performed if equation (4) is transformed into

$$(5) \quad c_1 \int_0^{M/\sigma_1} e^{-\frac{t^2}{2}} dt = c_2 \int_0^{\frac{b-M}{\sigma_2}} e^{-\frac{t^2}{2}} dt.$$

In distribution (1),  $\sigma_1$  and  $\sigma_2$  denote the standard deviations of the component distributions, measured from the means of the respective components. Hence, the standard deviation,  $\sigma$ , of the entire distribution satisfies the equation

$$\sigma = \sqrt{c_1(\sigma_1^2 + \bar{x}^2) + c_2(\sigma_2^2 + [b - \bar{x}]^2)}.$$

Therefore the value of the standard deviation of the arithmetic means of samples containing  $s$  items each drawn from distribution (1) is

$$(6) \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{s}} = \sqrt{\frac{c_1\sigma_1^2 + c_2\sigma_2^2 + c_1c_2b^2}{s}}.$$

If we assume that  $s$ , the number of items in the sample, is sufficiently large to justify its use, an approximation to the standard deviation of the median may be obtained from the equation

$$(7) \quad \sigma_M = \frac{1}{2y_M\sqrt{s}} \quad \text{where} \quad y_M = \frac{1}{\sqrt{2\pi}} \left( \frac{c_1}{\sigma_1} e^{-\frac{M^2}{2\sigma_1^2}} + \frac{c_2}{\sigma_2} e^{-\frac{(b-M)^2}{2\sigma_2^2}} \right).$$

## 2. The Relative Magnitude of the Median and Mean.

From equations (3) and (5) it is seen that, if four of the five parameters,  $c_1$ ,  $c_2$ ,  $\sigma_1$ ,  $\sigma_2$ ,  $b$ , are fixed and the fifth is allowed to vary, both  $\bar{x}$  and  $M$  will be monotone increasing functions of  $b$  and of  $c_2$ , and monotone decreasing functions of  $c_1$ , and that  $\bar{x}$  is independent of the standard deviations of both components, while  $M$  is a monotone increasing function of  $\sigma_1$  and a monotone decreasing function of  $\sigma_2$ .



When  $\mathcal{L}_1 = \mathcal{L}_2$  and  $\sigma_1 = \sigma_2$

distribution (1) becomes symmetrical, and

$$\bar{x} = M = \frac{\theta}{2}.$$

To obtain conditions under which  $\bar{x}$  shall exceed  $M$ , let equations (3) and (5) be differentiated with respect to  $\theta$ . The inequality

$$\frac{d\bar{x}}{d\theta} > \frac{dM}{d\theta}$$

may be reduced to the form

$$\frac{1}{\sigma_1} e^{-\frac{M^2}{2\sigma_1^2}} > \frac{1}{\sigma_2} e^{-\frac{(\theta-M)^2}{2\sigma_2^2}}.$$

It follows from equation (5) that when  $\mathcal{L}_1 > \mathcal{L}_2$ ,

$$\frac{\theta-M}{\sigma_2} > \frac{M}{\sigma_1}, \text{ whence } e^{-\frac{M^2}{2\sigma_1^2}} > e^{-\frac{(\theta-M)^2}{2\sigma_2^2}}.$$

Hence the inequalities

$$(8) \quad \mathcal{L}_1 > \mathcal{L}_2, \quad \sigma_2 \geq \sigma_1$$

are a sufficient condition that  $\frac{d\bar{x}}{d\theta}$  shall exceed  $\frac{dM}{d\theta}$ , and since  $\bar{x} = M = 0$  when  $\theta = 0$ , inequalities (8) are sufficient to insure that for positive values of  $\theta$ ,  $\bar{x}$  will exceed  $M$ .

In the case of many frequency distributions whose form suggests dissection into two normal components it is found that the standard deviations of the smaller component exceeds that of the larger component. Hence, condition (8) is fulfilled, and  $\bar{x}$  differs more from the mean of the larger component than does  $M$ .

### 3. Relative Stability of Median and Mean for the Special Case, $\theta=1$

From equations (6) and (7) it is seen that while  $\sigma_{\bar{x}}$  and  $\sigma_M$  are both monotone increasing functions of  $\theta$ , they do not possess a monotone character with respect to the other parameters of equation (1). The development of general conditions which the parameters must satisfy in order that  $\sigma_{\bar{x}}$  may exceed  $\sigma_M$  is impeded by

the fact that  $M$  is defined in (5) by an equation containing integrals, and its numerical value, for given values of the parameters, can be obtained only by interpolation in a table of areas under the normal curve. We shall therefore determine the relative stability of the median and arithmetic mean, as measured by the standard deviations of these averages, for certain special cases of distribution (1).

If, in equation (1),  $\theta$  is assigned the value zero, the distribution becomes symmetrical and  $\bar{x} = M = 0$ . Hence the condition for equal stability of median and arithmetic mean,  $\sigma_{\bar{x}} = \sigma_M$ , may in this special case be written

$$\sqrt{c_1 \sigma_1^2 + c_2 \sigma_2^2} = \sqrt{\frac{\pi}{2}} \frac{\sigma_1 \sigma_2}{c_1 \sigma_2 + c_2 \sigma_1}.$$

Letting the ratio,  $\frac{\sigma_2}{\sigma_1}$ , be denoted by  $\rho$ , we obtain

$$(9) \quad f(\rho) = c_1^2 c_2 \rho^4 + 2 c_1 c_2^2 \rho^3 + (c_1^3 + c_2^3 - \frac{\pi}{2}) \rho^2 + 2 c_1^2 c_2 \rho + c_1 c_2^2 = 0.$$

This fourth degree equation in  $\rho$  possesses two positive real roots, independent of  $c_1$ , and  $c_2$ , for  $f(0)$  and  $f(\infty)$  are both positive, while

$$f(1) = (c_1 + c_2)^3 - \frac{\pi}{2} = 1 - \frac{\pi}{2} < 0.$$

Hence there exist two values,  $\rho_1 < 1$  and  $\rho_2 > 1$ , such that when  $\rho$  assumes either of these values the standard deviations of the arithmetic mean and median are equal. For values of  $\rho$  in the interval  $(\rho_1 < \rho < \rho_2)$ , the standard deviation of the arithmetic mean is less than that of the median. For values of  $\rho$  outside this interval, the standard deviation of the arithmetic mean is greater than that of the median. Hence it is seen that, for  $\theta = 0$ , the relative stability of the median and arithmetic mean of distribution (1) is determined by the ratio of the standard deviations of the two component curves.

Yule<sup>8</sup> has discussed the relative stability of the median and arithmetic mean of distribution (1) when, in addition to the condition  $\theta = 0$ , the distribution is subjected to the further restriction

$$c_1 = c_2 = 0.5,$$

<sup>8</sup> Yule, G. U., *An Introduction to the Theory of Statistics*, 8th ed., p. 339.

and has obtained the numerical values of  $\rho$  for which the two averages will possess equal standard deviations:

$$\rho_1 = 0.4472, \quad \rho_2 = 2.2360.$$

4. *Relative Stability of Median and Mean for the Special Case,  $c_1 =$*

Now let the restriction  $\ell = 0$  be removed. Let  $\ell$  assume any positive value, and let the condition,  $c_1 = c_2 = 0.5$ , be imposed. The upper limits of the integrals in equation (5) will then be equal, whence

$$M = \frac{\ell \sigma_1}{\sigma_1 + \sigma_2} \qquad \bar{x} = \frac{\ell}{2}$$

$$\sigma_M = \frac{\sqrt{2\pi} \sigma_1 \sigma_2 e^{\frac{\ell^2}{2(\sigma_1 + \sigma_2)^2}}}{\sqrt{5} (\sigma_1 + \sigma_2)} \qquad \sigma_{\bar{x}} = \frac{\sqrt{2\sigma_1^2 + 2\sigma_2^2 + \ell^2}}{2\sqrt{5}}.$$

The relative magnitude of the median and mean is seen to depend upon the standard deviations of the component distributions, and  $\bar{x}$  is greater than, equal to, or less than  $M$  according as  $\rho = \sigma_2/\sigma_1$  is greater than, equal to, or less than unity.

To obtain conditions for equal stability in the two averages, let  $\sigma_{\bar{x}}$  be set equal to  $\sigma_M$ . By introducing the notation,

$$\rho = \sigma_2/\sigma_1, \quad k = \ell/(\sigma_1 + \sigma_2),$$

this equation may be reduced to the form

$$(10) \quad (k^2 + 2)\rho^4 + (4k^2 + 4)\rho^3 + (6k^2 + 4 - 8\pi e^{k^2})\rho^2 + (4k^2 + 4)\rho + (k^2 + 2) = 0.$$

Taking  $\lambda = (\rho + 1/\rho)$  as a new variable, this equation may be written as the quadratic

$$(k^2 + 2)\lambda^2 + (4k^2 + 4)\lambda + 4k^2 - 8\pi e^{k^2} = 0,$$

whose roots, are both real for all values of  $k^2$ . Furthermore, since  $\pi e^{k^2} > 2(k^2 + 1)$  for all values of  $k^2$ , one of these roots is positive and greater than 2, and therefore has a value which  $\lambda = (\rho + 1/\rho)$  may assume.

Hence, for all values of  $k^2$  (and therefore for all values of  $\ell$ ) there exist two reciprocal values of  $\rho$ , ( $\rho_1$  and  $\rho_2$ ), such

that when  $\rho$  assumes either of these values the standard deviations of the arithmetic mean and median are equal. For values of  $\rho$  in the interval  $(\rho_1 < \rho < \rho_2)$ , the standard deviation of the arithmetic mean is less than that of the median, and for values of  $\rho$  not in this interval, the standard deviation of the arithmetic mean is greater than that of the median.

Yule's results show that when  $t=0$ ,  $\rho_1=0.4472$  and  $\rho_2=2.2360$ , and therefore that the mean and median are equally stable when the standard deviation of one component is approximately 2.25 times that of the other. It remains to investigate the behavior of the interval  $(\rho_1 < \rho < \rho_2)$  as  $t$  varies.

Since it is the ratio of the standard deviations of the component curves, and not their actual values, which determines this interval, suppose the unit of measurement to be so chosen that  $(\sigma_1 + \sigma_2) = 1$ , whence  $t = t$ , and

$$\lambda = \frac{-2(t^2+1) + 2\sqrt{2\pi e^{t^2}(t^2+2)+1}}{t^2+2}.$$

Then since  $\pi e^{t^2}(t^2+2) > 2$  for all values of  $t^2$ ,  $\lambda$  may be shown to be a monotone increasing function of  $t^2$ , and therefore a monotone increasing function of  $t$  (positive). Since, furthermore,  $\lambda$  is an increasing function of  $\rho$  (when  $\rho > 1$ ), it appears that the interval  $(\rho_1 < \rho < \rho_2)$  in which the standard deviation of the arithmetic mean is less than that of the median (i. e., the interval in which the mean is the more stable average), becomes larger as the size of  $t$  is increased.

Summarizing, for the special case of distribution (1) in which  $\sigma_1 = \sigma_2$ , (i. e., in which the areas of the two component normal curves are equal), the relative stability of the median and mean depends upon the value of  $\rho$ , the ratio of the standard deviations of the two component curves, and upon the value of  $t$ , the distance between their means. When  $\rho$  equals one, the mean is the more stable average, independent of  $t$ . Furthermore, for all positive values of  $t$  there exists an interval of values of  $\rho$ , including  $\rho = 1$ , within which the mean is more stable, at the end points of which the averages are equally stable, and without which the median is more stable. When  $t=0$ , this interval is  $(.4472 < \rho < 2.2360)$ , and as  $t$  increases the interval becomes larger.

It was stated at the beginning of this section that, on account of the approximation to the value of  $\sigma_M$  which is used, the conclusions will apply to distributions containing a large number of items. It should be noted that, in the special case which has just been considered, to assign a large value to  $\theta$  may cause the median to fall at a point of relatively small frequency, in which case, as will be shown in section IV, the approximation to the standard deviation of the median,  $\sigma_M = 1/(2y_M\sqrt{5})$ , will exceed its true value, and the superior stability of the arithmetic mean, as obtained from equation (10), may be exaggerated. In such cases the probable errors of the averages should be computed by the method of section V.

### 5. Relative Stability of Median and Mean for the General Distribution (1).

Finally, let the restriction  $c_1 = c_2$  be removed from distribution (1). The condition that the standard deviations of the median and mean of this general distribution shall be equal may be written

$$\sqrt{c_1 \sigma_1^2 + c_2 \sigma_2^2 + c_1 c_2 \theta^2} = \sqrt{\frac{\pi}{2}} \frac{\sigma_1 \sigma_2}{c_1 \sigma_2 e^{-\frac{M^2}{2\sigma_1^2}} + c_2 \sigma_1 e^{-\frac{(\theta-M)^2}{2\sigma_2^2}}}.$$

Let the notation

$$\rho = \frac{\sigma_2}{\sigma_1}, \quad \theta = \kappa \sqrt{\sigma_1 \sigma_2}, \quad q = e^{-\frac{M^2}{2\sigma_1^2}}, \quad m = e^{-\frac{(\theta-M)^2}{2\sigma_2^2}}$$

be introduced. The parameters  $\rho, \kappa, q, m$  may thus assume only positive values, and  $q$  and  $m$  are not greater than unity. Then

$$f(\rho) = c_1^2 c_2 q^2 \rho^4 + (2c_1 c_2^2 q m + c_1^3 q^2 \kappa^4) \rho^3 + (c_2^3 m^2 + 2c_1^2 c_2 q m \kappa^2 + c_1^3 q^2 \frac{\pi}{2}) \rho^2 + (c_1 c_2^3 m^2 \kappa^2 + 2c_1^2 c_2 q m) \rho + c_1 c_2^2 m^2 = 0.$$

This equation may possess two real positive roots. As  $\rho$  approaches zero or positive infinity, it is seen that  $f(\rho)$  becomes positive, independent of  $c_1$  and  $c_2$ , and therefore that the standard deviation of the mean exceeds that of the median. If the equation possesses two distinct positive roots, there will be an interval of positive values of  $\rho$  for which the standard deviation of the median will exceed that of the mean. However, this interval does not necessarily contain the value,  $\rho = 1$ , as in the special case where

## III.

THE RELATIVE STABILITY OF THE MEDIAN AND ARITHMETIC  
MEAN OF A SYMMETRICAL FREQUENCY DISTRIBUTION WHICH  
IS COMPOSED OF THREE NORMAL DISTRIBUTIONS.

1. *The Relative Magnitude of the Standard Deviations of the  
Median and Mean.*

It will be the purpose of this section to investigate the relative stability of the median and arithmetic mean of a symmetrical frequency distribution which may be dissected into three normal distributions, two of which possess equal areas and equal standard deviations and whose means are translated equal distances to left and right, respectively, of the mean of the third distribution.

The equation of the frequency function which describes such a distribution is of the form

$$(1) \quad y = \frac{c_1}{\sigma_1 \sqrt{2\pi}} e^{-\frac{x^2}{2\sigma_1^2}} + \frac{c_2}{\sigma_2 \sqrt{2\pi}} \left( e^{-\frac{(x-\theta)^2}{2\sigma_2^2}} + e^{-\frac{(x+\theta)^2}{2\sigma_2^2}} \right),$$

where the areas of the components are connected by the relation

$$c_1 + 2c_2 = 1.$$

Since the distribution is symmetrical with respect to the  $y$ -axis, both the median and mean fall at the origin, and, if the approximation,  $\sigma_M = 1/(2y_m\sqrt{5})$ , is used, the standard deviations of the median and mean are readily expressed in terms of the parameters of equation (1):

$$\sigma_M = \frac{\sqrt{2\pi} \cdot \sigma_1 \sigma_2}{2\sqrt{5} \cdot (c_1 \sigma_2 + 2c_2 \sigma_1 e^{-\frac{\theta^2}{2\sigma_2^2}})},$$

$$\sigma_x = \frac{1}{\sqrt{5}} \cdot \sqrt{c_1 \sigma_1^2 + 2c_2 (\sigma_2^2 + \theta^2)}.$$

To obtain conditions under which the two averages will be equally stable, let the notation,

$$\rho = \frac{\sigma_2}{\sigma_1}, \quad n \sigma_2 = \theta,$$

be employed and let the standard deviations of median and mean be set equal to each other, whence we obtain the equation

$$\sqrt{1 + 2c_2[\rho^2(1 + \pi^2) - 1]} = \frac{\rho\sqrt{2\pi}}{2[\rho + 2c_2(e^{-\frac{\pi^2}{2}} - \rho)]},$$

which, if we let

$$(e^{-\frac{\pi^2}{2}} - \rho) = g \quad \text{and} \quad [\rho^2(1 + \pi^2) - 1] = h,$$

may be written

$$(2) \quad f(c_2) = 16g^2h c_2^3 + 8g(2\rho h + g)c_2^2 + 4\rho(\rho h + 2g)c_2 + \rho^2(2 - \pi) = 0.$$

Since  $c_1 + 2c_2 = 1$ , only the positive real roots of equation (2) which are less than 0.5 are of interest. Independent of the positive value assigned to  $\rho$  and  $\pi$ , this equation may have not more than two real roots in this interval, for both  $f(0)$  and  $f(0.5)$  are less than zero when  $\rho \neq 0$ .

If equation (2) possesses two real, distinct, positive roots less than  $c_2 = 1/2$ , then there will be a subinterval of the interval  $(0 < c_2 < 0.5)$  within which the standard deviation of the arithmetic mean will exceed the standard deviation of the median, at the end points of which the averages will be equally stable, and without which the standard deviation of the median will exceed that of the arithmetic mean. If the equation has no real roots in this interval, the standard deviation of the median will exceed that of the arithmetic mean throughout the interval.

The tangents to the curve whose equation is (2) are horizontal when  $c_2 = -\rho/2g$  and when  $c_2 = (-\rho h - 2g)/6gh$ . Since  $f(-\rho/2g)$  is negative for all values of  $\rho$  other than zero,  $-\rho/2g$  is a value of  $c_2$  for which the standard deviation of the median is greater than the standard deviation of the arithmetic mean. If there exists an interval of values of  $c_2$  for which the standard deviation of the arithmetic mean is greater than the standard deviation of the median, it will contain the value  $c_2 = (-\rho h - 2g)/6gh$ . Therefore the condition that such an interval exist is

$$0 < \frac{-\rho h - 2g}{6gh} < \frac{1}{2}, \quad f\left(\frac{-\rho h - 2g}{6gh}\right) > 0.$$

If the value,  $c_2 = (-\rho h - 2g)/6gh$  lies in the interval  $(0 < c_2 < 0.5)$  and if  $f(-\rho h - 2g)/6gh = 0$ , then equation (2) will possess a double root, and the standard deviation of the median will equal the standard deviation of the arithmetic mean for a single value of  $c_2$ ,  $c_2 = (-\rho h - 2g)/6gh$ , and will exceed it for all other values of  $c_2$ . The condition under which  $f(-\rho h - 2g)/6gh$  will vanish is that  $g/\rho h$  assume one of the values: 4.67284, -1.53327, -0.13957, for letting  $c_2 = (-\rho h - 2g)/6gh$ , equation (2) becomes

$$(3) \quad 8(g - \rho h)^3 - 27\pi\rho^2 h^2 g = 0,$$

which may be written as a cubic equation in  $g/\rho h$  whose roots have the above values.

## 2. *The Dissection of a Frequency Distribution into Three Normal Components.*

In order to apply the above conclusions in determining the relative stability of the averages of a particular sequence of economic data, it is necessary that the data be dissected into three normal distributions. A general method of determining the values of the five parameters of equation (1) from given frequency data will therefore be developed.

Karl Pearson<sup>\*</sup> has described a method for dissecting an asymmetrical frequency curve into two normal curves. He obtains expressions for the first five moments of the curve, which he solves, after lengthy algebraic manipulation, for the parameters. A similar procedure, the solution of moment equations, may be applied to a dissection into three normal curves. However, since the distribution has been assumed to be symmetrical, expressions for the odd moments vanish identically. Hence it is necessary to use moments as high as the eighth in order to obtain five equations from which the values of the parameters may be determined. While Pearson's method of setting up the moment equations may be used, his method of solution will not carry over to this case.

<sup>\*</sup> Loc. cit.



Given a frequency distribution of the variable  $x$  whose origin has been chosen at the arithmetic mean of the distribution, let  $M'_k$  denote the  $k^{\text{th}}$  moment of the distribution, and let  $M_k$  be set equal to the corresponding moment of the theoretical distribution whose equation is (1). We have, then, as the equations from which the five parameters of distribution (1) may be determined:

$$(4) \quad c_1 + 2c_2 = 1$$

$$c_1 \sigma_1^2 + 2c_2 (\sigma_2^2 + \sigma^2) = M_2$$

$$3c_1 \sigma_1^4 + 2c_2 (3\sigma_2^4 + 6\sigma^2 \sigma_2^2 + \sigma^4) = M_4$$

$$15c_1 \sigma_1^6 + 2c_2 (15\sigma_2^6 + 45\sigma^2 \sigma_2^4 + 15\sigma^4 \sigma_2^2 + \sigma^6) = M_6$$

$$105c_1 \sigma_1^8 + 2c_2 (105\sigma_2^8 + 420\sigma^2 \sigma_2^6 + 210\sigma^4 \sigma_2^4 + 28\sigma^6 \sigma_2^2 + \sigma^8) = M_8$$

Instead of carrying through the solution of these five equations, it has been found convenient to assign to  $\sigma_1$  a value equal to the standard deviation of a central group of items, and to retain only the first four moment equations to be solved for the other four parameters. Later the five equations will be used to correct this estimated value of  $\sigma_1$ .

If we let  $M'_k$  denote the  $k^{\text{th}}$  moment of the given distribution with  $\sigma_1$  as unit, denote  $\sigma/\sigma_1$  by  $u$ ,  $\sigma_2/\sigma_1$  by  $\rho$ , and eliminate  $c_2$  from these equations, we obtain

$$(5) \quad c_1 + (1-c_1)\rho^2(1+u^2) = M'_2$$

$$3c_1 + (1-c_1)\rho^4(3+6u^2+u^4) = M'_4$$

$$15c_1 + (1-c_1)\rho^6(15+45u^2+15u^4+u^6) = M'_6$$

Now eliminating  $c_1$ , this system of equations reduces to

$$3[M'_2 - \rho^2(1+u^2)] + (1-M'_2)\rho^4(3+6u^2+u^4) = M'_4[1 - \rho^2(1+u^2)]$$

$$15[M'_2 - \rho^2(1+u^2)] + (1-M'_2)\rho^6(15+45u^2+15u^4+u^6) = M'_6[1 - \rho^2(1+u^2)],$$

and when the notation  $\alpha = (1-M_2)$ ,  $\beta = (M_4' - 3)$ ,  $\gamma = (3M_2' - M_4') = -3\alpha - \beta$ ,

$$\delta = (M_6' - 15), \quad \epsilon = (15M_2' - M_6') = -15\alpha - \delta,$$

is introduced and the equations are written in descending powers of  $\rho^2$  they become

$$(6) \quad \begin{aligned} \rho^4 \alpha (3 + 6u^2 + u^4) + \rho^2 \beta (1 + u^2) + \gamma &= 0, \\ \rho^6 \alpha (15 + 45u^2 + 15u^4 + u^6) + \rho^2 \delta (1 + u^2) + \epsilon &= 0. \end{aligned}$$

Let Sylvester's method of elimination be applied to these two equations, making use of the property that the resultant,  $R$ , of the equations  $a_0 x^3 + a_1 x^2 + a_2 x + a_3 = 0$  and  $b_0 x^2 + b_1 x + b_2 = 0$  is

$$R = \begin{vmatrix} (a_0 b_2) & (a_1 b_2) - a_2 b_0 & -a_3 b_0 \\ (a_0 b_1) & (a_1 b_1) - a_2 b_2 & -a_3 b_1 \\ b_0 & b_1 & b_2 \end{vmatrix},$$

where  $(a_i b_j)$  denotes  $a_i b_j - a_j b_i$ .<sup>10</sup>

Let  $(1+u^2) = f_1(u^2)$ ,  $(3+6u^2+u^4) = f_2(u^2)$ ,  $(15+45u^2+15u^4+u^6) = f_3(u^2)$ .

Then

$$R = \begin{vmatrix} \alpha \gamma f_3 - \gamma \delta f_1 f_2 & -3\delta f_1^2 - \alpha \epsilon f_2 & -\beta \epsilon f_1 \\ \alpha \beta f_1 f_3 & \alpha \gamma f_3 - \alpha \delta f_1 f_2 & -\alpha \epsilon f_2 \\ \alpha f_2 & 3f_1 & \gamma \end{vmatrix} = 0,$$

and expanding and simplifying this determinant we obtain

$$\begin{aligned} (3\alpha\beta\gamma\epsilon - 2\gamma\delta\gamma^2)f_1 f_2 f_3 + (\alpha\gamma\delta^2 - \gamma\beta\delta\epsilon)f_1^2 f_2^2 \\ + (\beta^2\delta\gamma - \beta^3\epsilon)f_1^3 f_3 + \alpha\gamma^3 f_3^2 + \gamma^2\epsilon^2 f_2^3 = 0. \end{aligned}$$

Since  $\alpha, \beta, \gamma, \delta, \epsilon$  are constants, this equation may be written

$$A f_1 f_2 f_3 + B f_1^2 f_2^2 + C f_1^3 f_3 + D f_3^2 + E f_2^3 = 0.$$

If, finally,  $f_1, f_2$ , and  $f_3$  are replaced by their values as functions of  $u^2$ , this equation reduces to

$$\begin{aligned} (7) \quad u^{12}(A+B+C+D+E) + u^{10}(22A+14B+18C+30D+18E) \\ + u^8(159A+67B+93C+315D+117E) \\ + u^6(468A+132B+196C+1380D+324E) \\ + u^4(555A+123B+195C+2475D+351E) \\ + u^2(270A+54B+90C+1350D+162E) \\ + (45A+9B+15C+225D+27E) = 0, \end{aligned}$$

<sup>10</sup> Dickson, L. E., *First Course in Theory of Equations*, p. 150.

a sixth degree equation in  $u^2$  upon which the complete solution of the problem now turns, for having obtained a value of  $u^2$  from equation (7), the values of  $\rho^2$  and  $c$ , may be determined from equations (6) and (5), respectively. Since  $u = \ell/\sigma_2$ ,  $\rho = \sigma_2/\sigma_1$ , and  $c_1 + 2c_2 = 1$ , the parameters of equation (1) may be obtained.

It will be recalled that equation (7) has been obtained from the first four of equations (4), and that we have employed this equation to obtain values of  $c_1$ ,  $c_2$ ,  $\sigma_2$ ,  $\ell$  corresponding to an assigned value of  $\sigma_1$ . This estimated value of  $\sigma_1$  may be corrected, and corresponding corrections to the values of the other four parameters may be obtained, by use of the five equations (4).

Let equations (4) be written

$$f_i(c_1, c_2, \sigma_1, \sigma_2, \ell) = M_{2i-2} \quad (i = 1, 2, 3, 4, 5),$$

Let  $c'_1, c'_2, \sigma'_1, \ell'$  denote the values which the four parameters take on when  $\sigma_1$  is assigned the value  $\sigma'_1$ . Let  $\Delta c_1, \Delta c_2, \Delta \sigma_1, \Delta \ell$  denote the respective corrections which should be applied. Then using Taylor's theorem and neglecting terms which contain derivatives of higher order than the first, we obtain five linear equations in the five corrections:

$$\begin{aligned} f_i(c_1, c_2, \sigma_1, \sigma_2, \ell) &= f_i(c'_1 + \Delta c_1, c'_2 + \Delta c_2, \sigma'_1 + \Delta \sigma_1, \sigma'_2 + \Delta \sigma_2, \ell' + \Delta \ell) \\ &= f_i(c'_1, c'_2, \sigma'_1, \sigma'_2, \ell') + \Delta c_1 \frac{\partial f_i}{\partial c_1} + \Delta c_2 \frac{\partial f_i}{\partial c_2} + \Delta \sigma_1 \frac{\partial f_i}{\partial \sigma_1} + \Delta \sigma_2 \frac{\partial f_i}{\partial \sigma_2} + \Delta \ell \frac{\partial f_i}{\partial \ell} = M_{2i-2}. \end{aligned}$$

The corrected values of the parameters,  $c'_1 + \Delta c_1$ ,  $c'_2 + \Delta c_2$ ,  $\sigma'_1 + \Delta \sigma_1$ ,  $\sigma'_2 + \Delta \sigma_2$ ,  $\ell' + \Delta \ell$  may be regarded as second approximations to their true values, and further approximations may be obtained in the same fashion.

#### IV.

##### THE STANDARD DEVIATION OF THE MEDIANS OF SMALL SAMPLES.

##### 1. The Classical Approximation to the Standard Deviation of the Median.

In the preceding sections an approximation to the standard deviation of the median has been used, and the conclusions have

been assumed to be valid only when  $s$ , the number of items in the sample, is large. We wish, in the present section, to examine this approximation, and to compare the results which it produces with those obtained when other methods of determining the standard deviation of the median are employed.

The formula ordinarily used to compute the standard deviation of the medians of samples of  $s$  items each, drawn from a frequency distribution whose equation is  $y = f(x)$  and which satisfies the condition

$$(1) \quad \int_{-\infty}^{\infty} f(x) dx = 0.5 = \int_0^{\infty} f(x) dx$$

is

$$(2) \quad \sigma_M = \frac{1}{2 \cdot f(o) \cdot \sqrt{s}} \quad (11)$$

That this formula gives only an approximation to the true value of the standard deviation of the median and that the approximation may be rather poor for distributions of certain types is clear from the following derivation of the formula.

Let samples containing  $s$  items each be drawn from the distribution  $y = f(x)$  which satisfies condition (1). Let the proportion of items above  $x = o$  in each sample be denoted by  $(0.5 + d)$ . These observed values will tend to cluster around 0.5 as a mean, with a standard deviation of  $\frac{1}{2\sqrt{s}}$ . Let the deviation of the median of a sample from the median of the theoretical distribution,  $x = o$ , be denoted by  $e$ . Then if the number of items in the sample is sufficiently large to justify us in assuming that  $d$  is so small that we may regard the element of the frequency curve whose base is the interval  $(o, e)$ , and whose area is  $d$ , as approximately a rectangle, we may write

$$e = \frac{d}{f(o)}, \quad \text{whence} \quad \sigma_M = \frac{\sigma_d}{f(o)} = \frac{1}{2 \cdot f(o) \cdot \sqrt{s}}.$$

<sup>11</sup> Rietz, H. L., *Mathematical Statistics*, Carus Monograph III, p. 134.  
Yule, G. U., *Introduction to the Theory of Statistics*, 8th ed., p. 337.

This replacement of an element of a frequency curve by a rectangle can be justified only when  $e$ , the deviation of the median of the sample from the median of the theoretical distribution, is small. Hence there is reason to doubt whether formula (2) will give a close approximation to the value of the standard deviation of the median of samples which do not contain a large number of items. The formula would seem particularly untrustworthy when applied to a theoretical distribution in which the median falls at a point of relatively small frequency.

An expression for the standard deviation of the median which is not liable to the inaccuracies of approximation (2) may be derived as follows. Given the frequency function  $y = f(x)$  which satisfies condition (1), if a sample of  $(2n+1)$  items is drawn from this distribution, the probability that an item will fall in the interval  $(x, x+dx)$  approaches the limit  $y dx$  as  $dx$  approaches zero, the probability that an item will fall below  $x$  is  $\int_{-\infty}^x f(x) dx$ , and the probability that an item will fall above  $x$  is  $\int_x^{\infty} f(x) dx$ . Hence the limit, as  $dx$  approaches zero, of the probability that the median of the sample will fall in the interval  $(x, x+dx)$  is

$$(2n+1) C_n \left[ \int_{-\infty}^x f(x) dx \right]^n \cdot \left[ \int_x^{\infty} f(x) dx \right]^n f(x) dx,$$

and the square of the standard deviation of the median may be obtained from the equation

$$(3) \quad \sigma_M^2 = (2n+1) C_n \int_{-\infty}^{\infty} x^2 f(x) \left[ \int_{-\infty}^x f(x) dx \right]^n \cdot \left[ \int_x^{\infty} f(x) dx \right]^n dx.$$

The integrations involved in this equation may be difficult to perform unless  $f(x)$  is a simple function. Hence, we consider the rectangular distribution whose equations are

$$f(x) = 1, \quad \left(-\frac{1}{2} \leq x \leq \frac{1}{2}\right); \quad f(x) = 0, \quad \left(x < -\frac{1}{2}\right), \left(x > \frac{1}{2}\right),$$

and obtain

$$\sigma_M^2 = (2n+1) C_n \int_{-1/2}^{1/2} x^2 (0.25 - x^2)^n dx = \frac{1}{4(2n+3)}.$$

If we denote by  $\sigma_M'$  the approximation to the value of the standard deviation of the median obtained using formula (2),

we have for this distribution  $\sigma_M' = \frac{1}{2 \cdot f(\omega) \cdot \sqrt{2n+1}} = \frac{1}{2 \sqrt{2n+1}}$ ,

whence we have the relation

$$\sigma_M = \sigma_M' \sqrt{\frac{277+1}{2n+3}}.$$

It is observed that the approximation,  $\sigma_M'$ , exceeds the true value,  $\sigma_M$ , for all values of  $n$ , but that the error factor approaches unity as  $n$  increases, and is close to unity even for fairly small values of  $n$ .

## 2. *A General Method of Obtaining Upper and Lower Limits of the Standard Deviation of the Median.*

For distributions composed of two normal components the integrations involved in equation (3) can be performed only approximately, and this equation will serve only to determine upper and lower limits of the true value of the standard deviation of the median. A more straightforward method of obtaining these upper and lower limits, and one which is applicable to any frequency distribution, will be followed.

Let  $x_i$  denote the deviation of the  $i^{th}$  percentile of a distribution from the median, and let  $p_i$  denote the probability that the median of a sample of  $s$  items will fall between the  $i^{th}$  and  $(i+1)^{th}$  percentiles of the distribution from which the samples are drawn. Then a lower limit of the standard deviation of the medians of samples containing  $s$  items drawn from this distribution is given by the expression

$$\left[ \sum_{i=1}^{49} x_i^2 p_{i-1} + \sum_{i=51}^{99} x_i^2 p_i \right]^{1/2},$$

and an upper limit, by the expression

$$\left[ \sum_{i=0}^{49} x_i^2 p_i + \sum_{i=51}^{100} x_i^2 p_{i-1} \right]^{1/2},$$

where, in the case of a distribution in which the zeroth or hundredth percentile is at an infinite distance from the median,

$x_o$  denotes the largest value of  $x$  for which it is true that  $\int_{-\infty}^{x_o} f(x) dx < \epsilon$ , and  $x_{100}$  denotes the smallest value of  $x$  for which it is true that  $\int_{x_{100}}^{\infty} f(x) dx < \epsilon$ , where  $\epsilon$  is an arbitrarily small positive constant.

The values of  $x_i$  depend on the distribution, and are independent of the number of items in the sample. The values of  $p_i$  depend on the number of items in the sample and are independent of the form of the distribution. Approximations to the values of  $p_i$ , ( $i = 0, 1, 2, \dots, 99$ ), may be obtained by use of the DeMoivre-Laplace theorem<sup>12</sup>. In our notation the theorem may be stated:

The probability that  $m$  or more of the items of a sample containing  $(2m-1)$  items will fall to the right of the  $i$  <sup>th</sup> percentile of the distribution from which the sample is drawn is

$$P_i = \int_{\bar{x}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{\bar{x}^2}{2}} d\bar{x} \quad \text{where} \quad \bar{x} = \frac{m - (2m-1)(1-.01i) - 0.5}{\sqrt{(2m-1)(.01i)(1-.01i)}}$$

Then  $p_i = P_i - P_{i+1}$ .

Tables<sup>13</sup> of values of  $p_i$  for samples containing 7 and 51 items have been computed, and have been used in calculating upper and lower limits of the standard deviation of the medians of samples containing 7 and 51 items drawn from the distributions whose equations are

$$f_1(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

$$f_2(x) = \frac{1}{2\sqrt{2\pi}} \left( e^{-\frac{(x+2)^2}{2}} + e^{-\frac{(x-2)^2}{2}} \right),$$

$$f_3(x) = \frac{1}{\sqrt{2\pi}} \left( e^{-\frac{(\frac{4}{3}x)^2}{2}} + e^{-\frac{(4x)^2}{2}} \right).$$

<sup>12</sup> Rietz, H. L., *Mathematical Statistics*, Carus Monograph III, p. 35.

<sup>13</sup> These tables are included in the author's dissertation, which is filed in the library of the University of Wisconsin.

Approximations to the value of the standard deviation of the median have also been obtained using formula (2). The results are tabulated below.

STANDARD DEVIATION OF THE MEDIAN

	7 items.			51 items.		
	Upper Limit	Lower Limit	Formula (2)	Upper Limit	Lower Limit	Formula (2)
$f_1(x)$	0.4715	0.4462	0.4737	0.1838	0.1631	0.1755
$f_2(x)$	1.4701	1.4114	3.5003	0.8551	0.7733	1.2968
$f_3(x)$	0.2683	0.2521	0.2368	0.0938	0.0831	0.0877

We conclude that, when applied to samples containing a fairly small number of items, the results obtained using the customary formula for the standard deviation of the median may be very untrustworthy, particularly for a distribution in which the median falls at a point of relatively small frequency. We therefore shall propose another method for the comparison of the stability of the arithmetic mean and median, one which does not involve the computation of the standard deviation of these averages.

## V.

THE RELATIVE STABILITY OF THE MEDIAN AND ARITHMETIC MEAN, DETERMINED FROM THE FREQUENCY DISTRIBUTIONS OF THESE AVERAGES.

1. *The Frequency Distributions of the Median and Mean.*

Since the true value of the standard deviation of medians of samples containing items, drawn from a frequency distribution which is composed of two normal distributions, is not easily determinable, and since the customary approximation is not sufficiently accurate to justify its use in the study of small samples drawn from a distribution of this type, we shall develop a method of comparing the relative stability of the median and arithmetic mean, based not on the standard deviations of these two averages but on their frequency distributions.



Another consideration, aside from expediency, motivates the development of this method, for even if the standard deviations of the arithmetic mean and median could be accurately computed, they would not determine the relative stability of the two averages unless it is assumed that the frequencies of the mean and median are distributed in the same fashion. If, however, the equations of the frequency curves of the mean and median of samples of  $S$  items drawn from a given distribution are determined, then by comparing the deviations from the median of corresponding percentiles of these two averages, a judgment as to the relative stability of the two averages may be formed.

We shall assume the frequency curve of the arithmetic means of samples of  $(2n+1)$  items to be normal, independent of the form of the theoretical distribution from which the samples are drawn, and to possess a standard deviation of  $\sigma/\sqrt{2n+1}$ , where  $\sigma$  is the standard deviation of the theoretical distribution.<sup>14</sup> We proceed to determine the equation of the frequency curve of the medians of these samples.

Let the equation of the original distribution be  $y = f(x)$ , and let the condition

$$\int_{-\infty}^{\infty} f(x) dx = \frac{1}{2} = \int_0^{\infty} f(x) dx$$

be satisfied. Then the probability that the median of a sample of  $(2n+1)$  items will fall in the interval  $(x, x+dx)$  is the product of the probabilities that an item will fall in this interval and that of the remaining  $2n$  items,  $n$  will fall above this interval and  $n$  below this interval. We let  $y_M$  denote the frequency function according to which the medians of the samples are distributed, and obtain

$$\begin{aligned} y_M &= (2n+1) \cdot {}_{2n}C_n \cdot f(x) \left[ \int_{-\infty}^x f(x) dx \right]^n \left[ \int_x^{\infty} f(x) dx \right]^n dx \\ &= (2n+1) \cdot {}_{2n}C_n \cdot f(x) \left[ \frac{1}{2} + \int_0^x f(x) dx \right]^n \left[ \frac{1}{2} - \int_0^x f(x) dx \right]^n dx \\ (1) \quad &= (2n+1) \cdot {}_{2n}C_n \cdot f(x) \left\{ .25 - \left[ \int_0^x f(x) dx \right]^2 \right\}^n dx. \quad (15) \end{aligned}$$

<sup>14</sup> Rietz, H. L., *Mathematical Statistics*, Carus Monograph III, p. 127.

<sup>15</sup> A similar expression for the probability density of the median is

## 2. The Stability of an Average Determined from Its Probable Error.

Expressions for the frequency functions of the median and mean of samples containing  $(2n+1)$  items having been determined, we may form a judgement as to the relative stability of these two averages for a given distribution by comparing the deviations, from the median of the given distribution, of corresponding percentiles of the two averages. If some definite criterion of relative stability is desired, it seems natural to select the probable errors of the averages, where the term (probable error) is understood to have its original meaning, and not to denote a fixed multiple of the standard deviation of the average. We shall therefore proceed to determine the deviation, from the median, of a given percentile of the frequency distribution of medians of samples containing  $(2n+1)$  items drawn from the distribution whose equation is  $y=f(x)$ .

found in a paper by E. L. Dodd (Functions of Measurements under General Laws of Error, *Skandinavisk Aktuarietidskrift*, 1922, p. 150), and is there used in comparing the relative stability of the median and arithmetic mean of certain theoretical frequency distributions. However, the method used in Dodd's paper is to compare the probability densities of the two averages at the median of the original distribution, rather than to compare the deviations from the median of specific percentiles of the frequency curves of the two averages, as we shall do. Dodd uses Stirling's formula to obtain an approximation to the probability density at the median,

$$y_M(0) = \sqrt{\frac{2(2n+1)}{\pi}} f(0),$$

and represents the probability density of the arithmetic mean at the same point by the expression

$$y_{\bar{x}}(0) = \frac{\sqrt{2n+1}}{\sigma\sqrt{2\pi}},$$

where  $\sigma$  is the standard deviation of the original distribution.

It is readily seen that this method of comparison, when applied to small samples, would lead to exactly the same inaccuracies that would result if the relative stability of the two averages were determined by comparing their standard deviations, the customary approximation formula being used to obtain the value of the standard deviation of the median, since

$$\frac{y_M(0)}{y_{\bar{x}}(0)} = \sqrt{\frac{2(2n+1)}{\pi}} f(0) \div \frac{\sqrt{2n+1}}{\sigma\sqrt{2\pi}} = \frac{2\sqrt{2n+1} f(0)}{\frac{\sqrt{2n+1}}{\sigma}} = \frac{\sigma_{\bar{x}}}{\sigma_{M'}}.$$

Let  $S$  denote that fraction of the area under the frequency curve of the medians which is bounded by ordinates drawn to the curve at the points  $x=0$  and  $x=\ell$ . Our problem is to determine the value of  $\ell$  which corresponds to an assigned value of  $S$ , and from (1) the relationship between  $\ell$  and  $S$  is seen to be expressible in the form

$$(2) \quad S = (2n+1)_{2n} C_n \int_0^{\ell} f(x) \left\{ \frac{1}{4} - \left[ \int_0^x f(x) dx \right]^2 \right\}^n dx;$$

where  $S$  may be assigned any value in the interval ( $0 \leq S \leq 0.5$ ).

If the transformation

$$t = 2 \int_0^x f(x) dx$$

be applied to equation (2) it becomes

$$(3) \quad S = \frac{(2n+1)_{2n} C_n}{2^{2n+1}} \int_0^{\alpha} (1-t^2)^n dt, \text{ where } \alpha \text{ corresponds to } \ell,$$

Then

$$(4) \quad \frac{S 2^{2n+1}}{(2n+1)_{2n} C_n} = \frac{S \cdot 2^{2n+1} \cdot (n!)^2}{(2n+1)!} = \int_0^{\alpha} (1-t^2)^n dt$$

$$= \alpha - \frac{n}{3} \alpha^3 + \frac{n(n-1)}{2! \cdot 5} \alpha^5 - \frac{n(n-1)(n-2)}{3! \cdot 7} \alpha^7 + \dots \pm \frac{\alpha^{2n+1}}{2n+1},$$

and using Stirling's approximation, we obtain

$$(5) \quad \frac{2S\sqrt{\pi n}}{(2n+1)} = \int_0^{\alpha} (1-t^2)^n dt$$

$$= \alpha - \frac{n}{3} \alpha^3 + \frac{n(n-1)}{2! \cdot 5} \alpha^5 - \frac{n(n-1)(n-2)}{3! \cdot 7} \alpha^7 + \dots \pm \frac{\alpha^{2n+1}}{2n+1}.$$

It is observed from equation (3) that, for a fixed value of  $n$ ,  $\alpha$  is a monotone increasing function of  $S$ , and that  $\alpha=0$  when  $S=0$ , and  $\alpha=1$  when  $S=0.5$ . It is also observed that, for a fixed value of  $S$ ,  $\alpha$  is a monotone decreasing function of

$n$ , and that  $\lim_{n \rightarrow \infty} \alpha = 0$ .

Unless  $S$  is assigned a value near 0.5, we may obtain an approximation to the value of  $\alpha$  from equation (5) by neglecting terms containing powers of  $\alpha$  higher than the third. We wish to determine the degree of approximation which is introduced by dropping terms after the second from the second member of equation (5). To this end, we shall first ascertain an interval of values of  $S$  for which it is true that  $\alpha$  is not greater than the simple, decreasing function of  $n$ ,  $1/\sqrt{n}$ ; that is, we shall determine the interval of values of  $S$  which satisfy the inequality,

$$\frac{2S\sqrt{n}}{2n+1} \leq \frac{1}{\sqrt{n}} - \frac{n}{3} \frac{1}{n\sqrt{n}} + \frac{n(n-1)}{2!5} \frac{1}{n^2\sqrt{n}} - \frac{n(n-1)(n-2)}{3!7} \frac{1}{n^3\sqrt{n}} + \dots \pm \frac{1}{(2n+1)n^n\sqrt{n}},$$

or

$$S \leq \frac{2n+1}{2n\sqrt{n}} \left[ 1 - \frac{1}{3} + \frac{n-1}{2!5n} - \frac{(n-1)(n-2)}{3!7n^2} + \dots \pm \frac{1}{(2n+1)n^n} \right].$$

The second member of this inequality is greater than

$$\frac{1}{\sqrt{n}} \left[ 1 - \frac{1}{3} + \frac{n-1}{2!5n} - \frac{(n-1)(n-2)}{3!7n^2} + \dots \pm \frac{1}{(2n+1)n^n} \right],$$

and since the terms of the finite alternating series within parentheses obviously decrease in numerical value, their sum will exceed 2/3 for all positive values of  $n$ . Therefore the inequality

$$\alpha \leq \frac{1}{\sqrt{n}}$$

will certainly be satisfied for all values of  $S$  in the interval

$$|S| \leq \frac{2}{3\sqrt{n}} = 0.3761,$$

and therefore  $\alpha$  is not greater than  $1/\sqrt{n}$  when  $S$  is assigned a value corresponding to a percentile of the frequency distribution of the medians between the 13th and 87th percentiles. Certainly in determining the first and third quartiles of the frequency distribution of the medians,  $\alpha$  will be less than  $1/\sqrt{n}$ .

Since in equation (5) the value of  $S$  is given by a finite alternating series whose terms do not increase in numerical value,

the error involved in neglecting terms after the second will be less than the first term neglected:

$$\frac{2n+1}{2\sqrt{\pi n}} \frac{n(n-1)}{2!5} \alpha^5.$$

But, when  $|S| \leq 0.3761$ , it is true that

$$\begin{aligned} \frac{2n+1}{2\sqrt{\pi n}} \frac{n(n-1)}{2!5} \alpha^5 &\leq \frac{2n+1}{2\sqrt{\pi n}} \frac{n(n-1)}{2!5} \cdot \frac{1}{n^2\sqrt{n}} = \frac{2n^2-n-1}{20 n^2\sqrt{n}} \\ &= \frac{1}{10\sqrt{\pi}} \left(1 - \frac{1}{2n} - \frac{1}{2n^2}\right) < \frac{1}{10\sqrt{\pi}} = 0.0564. \end{aligned}$$

We conclude, therefore, that the value of  $\alpha$  obtained from equation (5) by neglecting powers of  $\alpha$  higher than the third corresponds to a percentile of the frequency distribution of the median which differs from the assigned value of  $S$  by not more than 0.05.

We see, then, that an approximation to the value,  $\ell$ , of a given percentile of the frequency distribution of the medians of samples containing  $(2n+1)$  items drawn from the theoretical distribution whose equation is  $y=f(x)$  may be obtained by solving for  $\alpha$  the third degree equation

$$(6) \quad \frac{2S\sqrt{\pi n}}{2n+1} = \alpha - \frac{\pi}{3} \alpha^3,$$

where

$$\alpha = 2 \int_0^{\ell} f(x) dx.$$

The tables of values of  $p_i$  mentioned in the preceding section afford a check on the accuracy of the results of equation (6) when  $(2n+1)$  is assigned the values 7 and 51. From these tables it is observed that the third quartile of the frequency distribution of the medians of samples containing 7 items falls at the 62nd percentile of the theoretical distribution from which the samples are drawn, and that for samples containing 51 items the third quartile of the medians falls near the 55th percentile of the original distribution. Hence, when  $S = 0.25$ , the value of  $\int_0^x f(x) dx$ , accurate to two places of decimals, is 0.12 when

( $2n+1$ ) equals 7, and 0.05 when ( $2n+1$ ) equals 51.

In equation (6) let

$$K = \frac{2.5 \sqrt{\pi n}}{2n+1},$$

whence we obtain

$$\frac{\pi}{3} \alpha^3 - \alpha + K = 0.$$

Letting  $\alpha = K + \lambda$ , this equation becomes

$$\frac{\pi}{3} (K^3 + 3K^2\lambda + 3K\lambda^2 + \lambda^3) - \lambda = 0,$$

or as an approximation

$$\lambda = \frac{\frac{\pi}{3} K^3}{1 - \pi K^2}.$$

Assigning to  $n$  the values 7 and 51 we obtain the following results:

$2n+1$	$K$	$\lambda$	$\alpha$	$\int_0^x f(x) dx$	Computed Value of $\int_0^x f(x) dx$
7	.2193	.0123	.2316	.1158	.12
51	.0869	.0072	.0941	.0471	.05

Thus we have developed a method of determining the probable error (or any percentile) of the median, which possesses the double advantage of being applicable to distributions which do not contain a very large number of items, and of being applied easily to any distribution, for after (6) has been used to determine the value of  $\alpha$ , the corresponding value of  $\epsilon$  may be obtained either from a table of integrals of a theoretical frequency function, or from an actual distribution by cumulating frequencies beyond the median.

The calculation of the probable error (or any percentile) of the arithmetic mean offers no difficulty if we assume the means of samples to be normally distributed. The relative stability of the two averages may then be determined by comparing the probable errors (or corresponding percentiles) of the two averages. This method of comparison will be applied to a particular distribution in section VI of this paper.

A table<sup>10</sup> of values of  $\alpha$  and  $\kappa$  for certain assigned values of  $n$  and  $S$  is given below.

Values of  $\alpha$  and  $\kappa$  when  $\kappa = \frac{2S\sqrt{\pi n}}{2\pi + 1} = \int_0^{\alpha} (1-t^2)^n dt$

$n$	$S$	$\kappa$	$\alpha$
10	.05	.02669	.027
	.10	.05338	.054
	.15	.08007	.082
	.20	.10676	.111
	.25	.13345	.143
	.30	.16014	.177
	.35	.18683	.217
	.40	.21352	.265
	.45	.24021	.334
25	.05	.017377	.017
	.10	.034754	.035
	.15	.052131	.053
	.20	.069508	.073
	.30	.104262	.116
	.35	.121639	.143
	.40	.139016	.176
	.45	.156393	.223
50	.05	.012409	.012
	.10	.024818	.025
	.15	.037227	.038
	.20	.049636	.052
	.25	.062045	.067
	.30	.074454	.083
	.35	.086863	.102
	.40	.099272	.126
	.45	.111681	.161
100	.05	.008818	.0088
	.10	.017636	.018
	.15	.026455	.027
	.20	.035273	.037
	.25	.044091	.047
	.30	.052909	.059
	.35	.061727	.073
	.40	.070546	.090
	.45	.079364	.115

<sup>10</sup> Computed by Miss Beatrice Berberich, university computer, University of Wisconsin.

## VI.

## AN APPLICATION TO A PARTICULAR SEQUENCE OF ECONOMIC DATA OF VARIOUS METHODS OF COMPARING THE STABILITY OF THE ARITHMETIC MEAN AND MEDIAN.

1. *Dissection into a Symmetrical Distribution Composed of Two Normal Distributions.*

In a paper by W. L. Crum<sup>17</sup> a particular sequence of economic data has been examined with the purpose of determining the relative stability of its median and arithmetic mean. The series studied comprises the monthly link relatives of the rate of interest on 60-90 day commercial paper from January, 1890, to January, 1917. A frequency distribution of deviations from their medians of the link relatives for each month is reproduced below, together with the values of the first six moments of the distribution.

FREQUENCIES OF DEVIATIONS FROM THE MEDIANS

Dev.	Freq.	Dev.	Freq.	Dev.	Freq.	Dev.	Freq.	Dev.	Freq.
-37	1	-18	2	-7	6	4	19	15	1
-32	1	-17	2	-6	23	5	13	16	1
-30	1	-16	2	-5	10	6	13	17	1
-29	1	-15	1	-4	13	7	8	18	2
-28	1	-14	3	-3	19	8	6	23	1
-24	1	-13	6	-2	9	9	5	24	1
-23	1	-12	3	-1	11	10	2	28	1
-22	1	-11	6	0	28	11	4	34	1
-21	2	-10	3	1	22	12	3	35	2
-20	1	-9	5	2	22	13	1	41	2
-19	2	-8	11	3	13	14	2	42	1
								45	1

$$N = 324$$

Moments	About 0 Deviation	About $\bar{x}$	With Sheppard Adjustments
1	-0.46	0.00	0.00
2	107.06	106.85	106.77
3	656.45	793.68	793.68
4	83520	84860	84465
5	3015000	3209000	3208000
6	110890000	119480000	119370000

<sup>17</sup> Loc. cit.



Professor Crum's method of attack is to dissect the series, according to Pearson's method, into two normal components whose means are coincident. He therefore fits to the data a curve whose equation is

$$(1) \quad y = \frac{324}{\sqrt{2\pi}} \left( \frac{c_1}{\sigma_1} e^{-\frac{x^2}{2\sigma_1^2}} + \frac{c_2}{\sigma_2} e^{-\frac{x^2}{2\sigma_2^2}} \right),$$

and obtains for the parameters the values:

$$c_1 = .26, \quad \sigma_1 = 2.46, \quad c_2 = .74, \quad \sigma_2 = 11.9.$$

This theoretical distribution is of the type discussed in paragraph 3, section II. Its median and mean will be equally stable if  $\rho = \sigma_2/\sigma_1$  satisfies the equation

$$c_1^2 c_2 \rho^4 + 2c_1 c_2^2 \rho^3 + (c_1^3 + c_2^3 - \frac{\pi}{2}) \rho^2 + 2c_1^2 c_2 \rho + c_1 c_2^2 = 0.$$

Letting  $c_1 = 0.25$  and  $c_2 = 0.75$ , this equation reduces to

$$\rho^4 + 6\rho^3 - 24\rho^2 + 2\rho + 3 = 0,$$

which has a root between 2.5 and 2.6. Since for the distribution under consideration  $\rho = 4.8$ , the standard deviation of the arithmetic mean is larger than the standard deviation of the median, and the median is the more stable average.

## 2. Dissection into an Asymmetrical Distribution Composed of Two Normal Distributions.

In the method of dissection employed by Professor Crum, the slight positive skewness which the distribution possesses is ignored. We shall dissect the data into two normal components whose means are not equal, and investigate the relative stability of the median and mean of the resulting asymmetrical distribution:

$$(2) \quad y = \frac{324}{\sqrt{2\pi}} \left[ \frac{c_1}{\sigma_1} e^{-\frac{(x-\xi_1)^2}{2\sigma_1^2}} + \frac{c_2}{\sigma_2} e^{-\frac{(x-\xi_2)^2}{2\sigma_2^2}} \right].$$

Pearson's method of dissecting an asymmetrical distribution depends on the solution of his "fundamental nonic,"

$$24\mu_2^9 - 28\lambda_4\mu_2^7 + 36\mu_3^2\mu_2^6 - (24\mu_3\lambda_5 - 10\lambda_4^2)\mu_2^5 - (148\mu_3^2\lambda_4 + 2\lambda_5^2)\mu_2^4 + (288\mu_3^4 - 12\lambda_4\lambda_5\mu_3 - \lambda_4^3)\mu_2^3 + (24\mu_3^3\lambda_5 - 7\mu_3^2\lambda_4^2)\mu_2^2 + 32\mu_3^4\lambda_4\mu_2 - 24\mu_3^6 = 0.$$

in which  $\mu_i$  denotes the  $i^{\text{th}}$  moment of the given distribution,

and  $\lambda_4 = 9\mu_2^2 - 3\mu_4$ ,  $\lambda_5 = 30\mu_2\mu_3 - 3\mu_5$ .

A value of  $\rho_2$  having been obtained from this equation, the parameters of equation (2) are determined by solving, successively, the equations

$$\rho_3 = \frac{2\mu_3^3 - 2\mu_3\lambda_4\rho_2 - \lambda_5\rho_2^2 - 8\mu_3\rho_2^3}{4\mu_3^2 - \lambda_4\rho_2 + 2\rho_2^3},$$

$$\rho_1 = \rho_3/\rho_2,$$

$\theta^2 - \rho_1\theta + \rho_2 = 0$ , (the two roots of this equation are denoted  $\theta_1, \theta_2$ )

$$c_1 = -\theta_2/(\theta_1 - \theta_2),$$

$$c_2 = \theta_1/(\theta_1 - \theta_2),$$

$$\sigma_1^2 = \mu_2 - \frac{1}{3} \frac{\mu_3}{\theta_1} - \frac{1}{3} \rho_1 \theta_1 + \rho_2,$$

$$\sigma_2^2 = \mu_2 - \frac{1}{3} \frac{\mu_3}{\theta_2} - \frac{1}{3} \rho_1 \theta_2 + \rho_2.$$

The calculation of the Sturm's functions of the fundamental nonic shows it to have three real roots, two between 0 and -100, and a third between 200 and 300. The values of these roots are found to be

$$\rho_2 = -5.5517, \quad -11.6140, \quad 210.$$

However, the use of the second and third of these roots leads to imaginary values of certain of the parameters of equation (2), and they are therefore rejected. Using the root,  $\rho_2 = -5.5517$ , the parameters of equation (2) are found to have the following values:

$$c_1 = 0.9637, \quad \theta_1 = -0.46, \quad \sigma_1 = 9.02$$

$$c_2 = 0.0363, \quad \theta_2 = 12.20, \quad \sigma_2 = 25.07.$$

### 3. *Dissection into a Symmetrical Distribution Composed of Three Normal Distributions.*

Finally, let the given data be fitted by a frequency curve whose equation is

$$(3) \quad y = \frac{324}{\sqrt{2\pi}} \left[ \frac{c_1}{\sigma_1} e^{-\frac{x^2}{2\sigma_1^2}} + \frac{c_2}{\sigma_2} \left( e^{-\frac{(x-b)^2}{2\sigma_2^2}} + e^{-\frac{(x+b)^2}{2\sigma_2^2}} \right) \right],$$

where the origin is selected at the arithmetic mean of the original series. The dissection depends on the solution of equation (7), section III, which for the distribution under consideration has the form

$$u^{12} - 58.355 u^{10} - 230.538 u^8 - 243.922 u^6 - 60.184 u^4 - 3.164 u^2 - 0.527 = 0.$$

The only positive root of this equation is  $u^2 = 62.13$ .

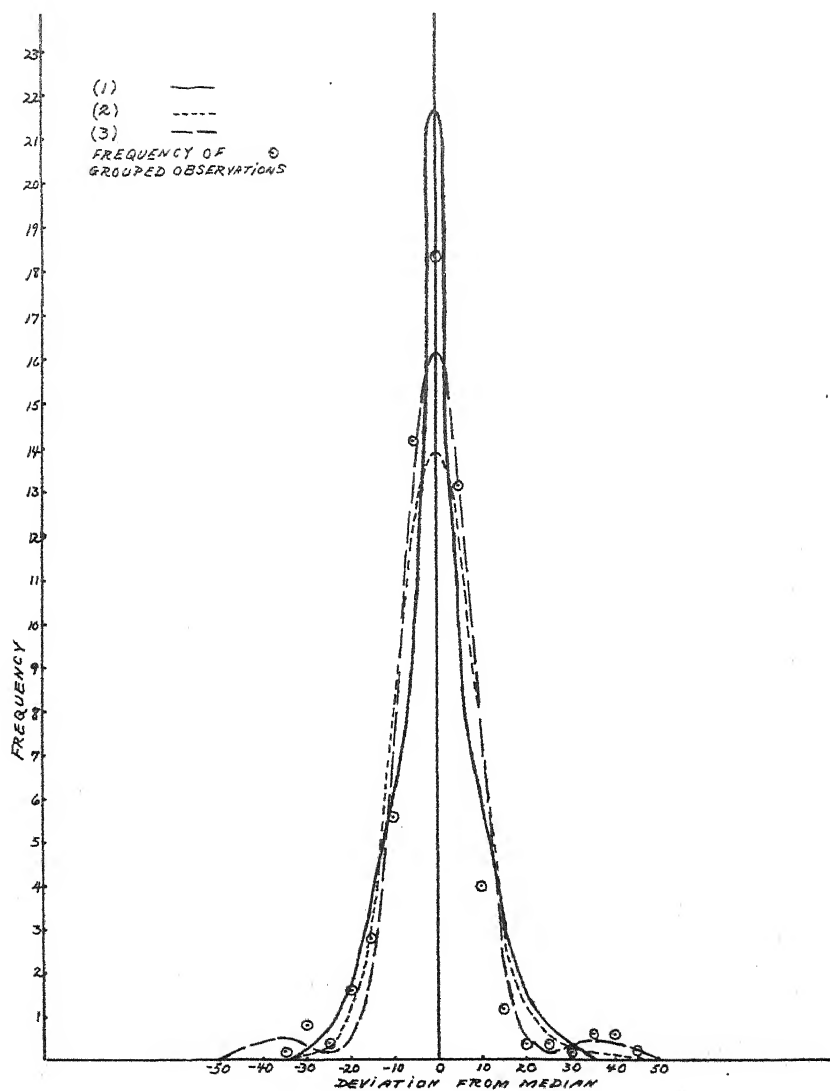
Solving, successively, equations (6) and (5) of section III, the values of the parameters of equation (3) are found to be

$$c_1 = 0.96, \quad \sigma_1 = 7.64, \quad c_2 = 0.02, \quad \sigma_2 = 4.69, \quad b = 36.95.$$

The accompanying figure shows the original distribution (grouped into class intervals of five units) and the curves obtained by each of the three methods of dissection, plotted on the same set of axes. It appears from the figure that a distribution of type (3) fits the data more closely than either of the other curves. This fact may be checked by comparing the sums of the squares of the differences between the actual and theoretical frequency of each class. The values of these sums of squares of deviations from theoretical distributions (1), (2), (3) are found to be, respectively, 68.250, 51.604, 19.435.

### 4.—*Relative Stability of Median and Mean for Each of the Methods of Dissection.*

We turn now to the problem of comparing the stability of the median and mean of the three theoretical frequency functions



obtained by dissection. Since  $N=324$  is fairly large, and since the median of each of the theoretical distributions is located at a point of relatively large frequency, we shall use the approximation to the standard deviation of the median,

$$\sigma_M = \frac{1}{2 \cdot y_M \cdot \sqrt{5}},$$

where  $y_M$  is the ordinate at the median.

For Crum's dissection into two normal distributions, the arithmetic mean and median are both at the origin. Hence

$$\sigma_{\bar{x}} = \frac{1}{\sqrt{324}} \sqrt{c_1 \sigma_1^2 + c_2 \sigma_2^2} = 0.57,$$

$$\sigma_M = \frac{1}{\frac{2 \sqrt{324}}{\sqrt{2\pi}} \left( \frac{c_1}{\sigma_1} + \frac{c_2}{\sigma_2} \right)} = 0.42,$$

which verifies his conclusion that the median is more stable than the arithmetic mean.

For the asymmetrical dissection into two normal curves,

$$\bar{x} = c_2 b = 0.46$$

$$\sigma_{\bar{x}} = \frac{1}{\sqrt{324}} \sqrt{c_1 \sigma_1^2 + c_2 \sigma_2^2 + c_1 c_2 b^2} = 0.57.$$

$M$  is a root of the equation

$$c_1 \int_0^{\frac{M}{\sigma_1}} e^{-\frac{t^2}{2}} dt = c_2 \int_0^{\frac{b-M}{\sigma_2}} e^{-\frac{t^2}{2}} dt,$$

and its value, obtained by interpolation in a table of values of the integral  $\int_0^x e^{-\frac{t^2}{2}} dt$ , is  $M=0.16$ . Hence

$$\sigma_M = \frac{1}{\frac{2 \sqrt{324}}{\sqrt{2\pi}} \left( \frac{c_1}{\sigma_1} e^{-\frac{M^2}{2\sigma_1^2}} + \frac{c_2}{\sigma_2} e^{-\frac{(b-M)^2}{2\sigma_2^2}} \right)} = 0.64.$$

For this method of dissection the arithmetic mean is more stable than the median. This was to be expected, since the second component contains so small a fraction of the total area that the compound curve differs little from a single normal curve. The

figure shows that this curve would not naturally be chosen to represent the given data.

For the symmetrical three normal curve dissection,

$$M = \bar{x} = 0,$$

$$\sigma_{\bar{x}} = \frac{1}{\sqrt{324}} \sqrt{c_1 \sigma_1^2 + 2c_2 \sigma_2^2 + 2c_2 b^2} = 0.59,$$

$$\sigma_M = \frac{1}{\sqrt{324}} \frac{\sqrt{2\pi} \cdot \sigma_1 \sigma_2}{2(c_1 \sigma_2 + 2c_2 \sigma_1 e^{-b^2/2\sigma_2^2})} = 0.55.$$

This method of dissection bears out Crum's conclusion that the median of the original series is a more stable average than the arithmetic mean, although the difference between the standard deviations of the two averages obtained by this method is considerably smaller than that obtained by the first method of dissection.

#### 5. *The Probable Errors of the Median and Mean, Determined from the Frequency Distributions of these Averages.*

The above discussion has been concerned with certain theoretical frequency curves, rather than with the actual data which these curves are intended to fit. We shall now compare the relative stability of the mean and median by the method of section V, which does not involve a fitting to the data of a theoretical frequency curve.

The method of determining the quartiles of the frequency distribution of the median, developed in section V, assumed the number of items in the sample to be odd. We therefore solve the equations

$$\kappa = \frac{\frac{1}{2} \sqrt{\pi n}}{2\pi + 1}, \quad \lambda = \frac{\frac{\pi}{3} \kappa^3}{1 - \pi \kappa^2},$$

using the values 323 and 325 for  $(2\pi + 1)$  and obtain roots

$$\kappa = 0.0348, \quad \lambda = 0.0027$$

in each case, whence  $\alpha = \kappa + \lambda = 0.0375$ ,

and 
$$\int_0^x f(x) dx = 0.0188.$$

For the given distribution, the median falls at zero deviation. The fourteen items in the upper half of the zero class comprise 4.32% of the entire frequency distribution. Hence the third quartile of the distribution of the medians has a value

$$Q = \frac{1}{2} \cdot \frac{0.0188}{0.0432} = 0.2176.$$

Similar reasoning shows the value of the first quartile of the medians to be—0.2176, whence the semi-interquartile range is 0.2176.

Since  $\sigma^2 = 106.85$ , the probable error of the arithmetic mean has a value

$$.6745 \sqrt{\frac{106.85}{324}} = 0.3845.$$

Thus the median is again shown to be more stable than the arithmetic mean.

HARRY S. POLLARD,  
Miami University,  
Oxford, Ohio.

*Harry S. Pollard*

# AN APPLICATION OF CHARACTERISTIC FUNCTIONS TO THE DISTRIBUTION PROBLEM OF STATISTICS\*

By

SOLOMON KULLBACK,

*George Washington University, Washington, D. C.*

## CONTENTS

	PART I	SECTION
Introduction .....		I
Characteristic Functions .....		II
Theorems Regarding a Single Function, $\mu(x_1, \dots, x_n)$ .		III
Theorems Regarding Several Functions, $\mu(x_1, \dots, x_n), j=1, 2, \dots, n$ .		IV
PART II		
Distribution of the Arithmetic Mean .....		V
Distribution of the Geometric Mean .....		VI
Lemma .....		VII
Distribution of Variance of a Sample of $n$ From a		
Normal Population .....		VIII
Distribution of the $\chi^2$ of Goodness of Fit Test.....		IX
Simultaneous Distribution of Variances and Correlation		
Coefficient of a Sample of $n$ from a Bi-variate Normal Population .....		X
Distribution of the Covariance of a Sample of $n$ from		
a Bi-variate Normal Population .....		XI
Do $N$ Samples of $n$ -categories, come from the Same		
$n$ -variate Normal Population? .....		XII
Distribution of the Generalized Variance of a Sample of		
$N$ from an $n$ -variate Normal Population .....		XIII
PART III		
Summary and Conclusions .....		XIV

---

\* Presented to the American Mathematical Society, under the title, "An Application of Characteristic Functions to Statistics," Feb. 25, 1933. This paper was prepared under the guidance of Professor F. M. Weida.



## PART 1

*The General Theory*

*I. Introduction:*\* By the distribution problem of statistics we mean the problem of determining the distribution law of functions of variables satisfying known distribution laws. Many particular problems of this nature have been solved by various methods. In Part 1 of this paper we develop a general solution for this problem for functions of variables satisfying continuous distribution laws. The general result is then applied in Part 2 to derive the distribution laws of several functions whose distribution laws have been derived by other methods and of some functions whose distribution laws have not been given or given only for special cases; in Part 3 we summarize the results. The method of solution is related to the concept of characteristic function.

The theory of characteristic functions is essentially a development of Laplace's<sup>18</sup> "fonction génératrice." In this paper we shall adopt the term characteristic function, although the same concept has been termed generating function<sup>14</sup> and reciprocal function.<sup>21</sup> Poisson<sup>28, 29</sup> employed the methods of Laplace to discuss, in particular, "Sur la Probabilité des Resultats Moyens des Observations." Cauchy<sup>2</sup> was apparently the next to study and apply this theory; he applied the basic concept of characteristic function in connection with what he called "coefficient limitateur ou ristricteur" to study the problem of a function of errors. In particular he studied the case of a linear function of the errors. More recently the same concept has been reintroduced under the name of characteristic function by Poincaré<sup>27</sup> and also by P. Lévy<sup>17, 18, 19</sup> who employs it to consider the composition of laws of probability, the notion of the limit of a probability law, the idea of stable and semi-stable laws, etc.

In a series of papers, C. V. L. Charlier<sup>3</sup> further applied and

---

\* The reference numbers correspond with the number of the item in the bibliography.

developed the theory of characteristic functions (though he employed the terminology of reciprocal functions) to develop the Gram-Charlier Type A and Type B series, and to consider the distribution law of functions of variables satisfying general frequency laws. Under the name of "Erzeugenden Funktion," T. Kameda<sup>14</sup> studied the properties of functions which are intimately related to characteristic functions. In particular, he discussed the development of a function as a series of Hermite Polynomials and also considered the problem of finding the distribution law of a function of variables obeying general distribution laws.<sup>15</sup>

*II. Characteristic Functions:* By the characteristic function of the distribution law of the variable  $x$  is meant the mean\* value of  $e^{itx}$  where  $i = \sqrt{-1}$ . Thus, for a continuous distribution law, if  $f(x)dx$  is the probability to within infinitesimals of a higher order that  $x - \frac{dx}{2} < x, < x + \frac{dx}{2}$  and  $\varphi(t)$  is the characteristic function of the d.l.,\*\* of  $x$  then

$$(1) \quad \varphi(t) = \int e^{itx} \cdot f(x) dx,$$

where the limits of the integral depend upon the range of applicability of  $f(x)$ . We may also write

$$\varphi(t) = \int_{-\infty}^{\infty} e^{itx} f(x) dx \text{ if we}$$

agree that  $f(x) \equiv 0$  outside the range of applicability. The characteristic function derives its importance from the fact<sup>17</sup> that

$$(2) \quad f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \varphi(t) dt.$$

For the case of several variables, we have that the character-

\* Also known as probable or expected value.

\*\* We shall designate distribution law hereafter by d.l.

istic function of the d.l.  $f(x_1, x_2, \dots, x_n)$  of  $x_1, x_2, \dots, x_n$  is given by

$$(3) \quad \varphi(t_1, t_2, \dots, t_n) = \int \cdots \int_R e^{it_1 x_1 + it_2 x_2 + \cdots + it_n x_n} f(x_1, \dots, x_n) dx_1 \cdots dx_n,$$

where  $R$  is the region of applicability of  $f(x_1, x_2, \dots, x_n)$ . We

$$\text{may also write } \varphi(t_1, \dots, t_n) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{it_1 x_1 + \cdots + it_n x_n} f(x_1, \dots, x_n) dx_1 \cdots dx_n$$

provided we agree that  $f(x_1, x_2, \dots, x_n) \equiv 0$  outside the region  $R$ . As for the case of a single variable we have here too<sup>88</sup>

$$(4) \quad f(x_1, x_2, \dots, x_n) = \frac{1}{(2\pi)^n} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-it_1 x_1 - \cdots - it_n x_n} \varphi(t_1, \dots, t_n) dt_1 \cdots dt_n.$$

We shall prove that the following extensions are also possible. Consider the function  $u(x_1, x_2, \dots, x_n)$  \* of the variables  $x_1, x_2, \dots, x_n$  whose d.l. is  $f(x_1, x_2, \dots, x_n)$ . Then the characteristic function of the d.l. of  $u$  is given by

$$(5) \quad \varphi(t) = \int \cdots \int_R e^{it \cdot u(x_1, x_2, \dots, x_n)} f(x_1, x_2, \dots, x_n) dx_1 \cdots dx_n,$$

where  $R$  is the region of applicability of  $f(x_1, x_2, \dots, x_n)$ . The d.l. of  $u$ ,  $P(u)$ , is given by

$$(6) \quad P(u) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-it u} \varphi(t) dt, \quad \text{where}$$

$\varphi(t)$  is defined by (5).

If we consider the several functions  $u_1(x_1, x_2, \dots, x_n)$  ;

---

\* The conditions which  $u(x_1, x_2, \dots, x_n)$  must satisfy will be developed further in this paper.

$u_2(x_1, x_2, \dots, x_n); \dots; u_n(x_1, x_2, \dots, x_n)$  of the variables  $x_1, x_2, \dots, x_n$  whose d.l. is  $f(x_1, x_2, \dots, x_n)$ , then the characteristic function of the d.l. of  $u_1, u_2, \dots, u_n$  is given by

$$(7) \quad \varphi(t_1, t_2, \dots, t_n) = \int_R \int e^{it_1 u_1(x_1, x_2, \dots, x_n) + \dots + it_n u_n(x_1, x_2, \dots, x_n)} \cdot f(x_1, x_2, \dots, x_n) dx_1, dx_2, \dots, dx_n,$$

where  $R$  is the region of applicability of  $f(x_1, x_2, \dots, x_n)$ . The d.l. of  $u_1, u_2, \dots, u_n$  is given by

$$(8) \quad P(u_1, u_2, \dots, u_n) = \frac{1}{(2\pi)^n} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-it_1 u_1 - it_2 u_2 - \dots - it_n u_n} \cdot \varphi(t_1, t_2, \dots, t_n) dt_1 dt_2 \dots dt_n$$

where  $\varphi(t_1, t_2, \dots, t_n)$  is defined by (7).

III. *Theorems Regarding a Single Function*  $u(x_1, x_2, \dots, x_n)$ : We shall now justify our statements and determine the precise conditions the function  $u$  must obey.

Consider the function  $u(x_1, x_2, \dots, x_n)$  of the variables  $x_1, x_2, \dots, x_n$  satisfying the continuous d.l.  $f(x_1, x_2, \dots, x_n)$

such that  $\int_R \dots \int f(x_1, x_2, \dots, x_n) dx_1 \cdot dx_2 \dots dx_n = 1$ . The

function  $u$  may have at most a denumerable infinity of discontinuities. The probability that  $u(x_1, x_2, \dots, x_n)$  satisfies the conditions

$$(9) \quad u_1 < u < u_2 \quad \text{is given by}$$

$$(10) \quad \int_A \dots \int f(x_1, x_2, \dots, x_n) dx_1 \cdot dx_2 \dots dx_n, \text{ where } A$$

is the region defined by the inequalities  $u_1 < u < u_2$ . To avoid the difficulty of integrating over the region  $A$  we shall avail ourselves of the discontinuity factor (See Whittaker and Watson<sup>40</sup> § 9.7).

$$(11) \quad F = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{u_1}^{u_2} e^{-it(\theta-u)} d\theta dt, \quad \text{where } F = 1$$

for  $u_1 < u < u_2$ ;  $F = 0$  for  $u \geq u_1$ ;  $F = 0$  for  $u \geq u_2$ .

We are now able to say that the required probability is given\* by

$$(12) \quad \int_R \cdots \int f(x_1, x_2, \dots, x_n) \cdot F \cdot dx_1 dx_2 \cdots dx_n.$$

If we set  $2\omega = u_1 + u_2$  and  $\gamma = u_2 - u_1$ , the required probability may also be written as

$$(13) \quad \frac{1}{2\pi} \int_R \cdots \int f(x_1, x_2, \dots, x_n) dx_1 dx_2 \cdots dx_n \cdot \int_{-\infty}^{\infty} \int_{u-\frac{\gamma}{2}}^{u+\frac{\gamma}{2}} e^{-it(\theta-u)} d\theta dt.$$

Integrating with respect to  $\theta$ , we obtain

$$(14) \quad \frac{1}{2\pi} \int_R \cdots \int f(x_1, \dots, x_n) dx_1 \cdots dx_n \cdot \int_{-\infty}^{\infty} e^{it(u-\omega)} \cdot \frac{2 \sin \frac{\gamma t}{2}}{t} dt.$$

We now want to prove that

$$(15) \quad \int_R z d\mathbf{X} \int_{-\infty}^{\infty} e^{it(u-\omega)} \cdot \frac{2 \sin \frac{\gamma t}{2}}{t} dt = \int_{-\infty}^{\infty} \frac{2 \sin \frac{\gamma t}{2}}{t} dt \int_R e^{it(u-\omega)} z d\mathbf{X},$$

where we write  $z = f(x_1, x_2, \dots, x_n)$ ;  $d\mathbf{X} = dx_1 dx_2 \cdots dx_n$

\* This method is essentially an application of Cauchy's "Coefficient limitateur ou ristricteur." See C.R. Vol. 37, p. 150 ff, and Whittaker and Robinson, *Calculus of Obs.*, p. 169.

and  $\int_R$  as the multiple integral over the region  $R$ .

We have that

$$(16) \quad \int_R z dX \int_{-\infty}^{\infty} \frac{2 \sin \frac{\alpha t}{2}}{t} e^{it(u-w)} dt = \int_R z dX \int_0^{\infty} \frac{2 \sin \frac{\alpha t}{2}}{t} e^{it(u-w)} dt \\ + \int_R z dX \int_0^{\infty} \frac{2 \sin \frac{\alpha t}{2}}{t} e^{-it(u-w)} dt.$$

We will now prove that

$$(17) \quad \int_R z dX \int_0^{\infty} \frac{2 \sin \frac{\alpha t}{2}}{t} e^{it(u-w)} dt = \int_0^{\infty} \frac{2 \sin \frac{\alpha t}{2}}{t} dt \int_R e^{it(u-w)} \cdot z \cdot dX$$

For this, it is sufficient<sup>12</sup> to prove the existence of the  $(n+1)$  fold integral\*  $\int z \frac{2 \sin \frac{\alpha t}{2}}{t} e^{it(u-w)} dX dt$ ,

and the existence of the right-hand member of (17).

Consider\*\* the rectangular region  $G$  in  $(n+1)$  fold space defined by  $0 \leq t \leq t_1$ ;  $x'_j \leq x_j \leq x''_j$ ,  $j = 1, 2, 3, \dots, n$ , where we shall designate the region  $x'_j \leq x_j \leq x''_j$ ,  $j = 1, 2, \dots, n$ , by  $E$ . Then, over  $G$  the multiple integral of

$$z \cdot \frac{2 \sin \frac{\alpha t}{2}}{t} e^{it(u-w)}$$

exists since the integrand is bounded and has at most a denumerable infinity of singularities (those of  $u(x_1, x_2, \dots, x_n)$ ). Then<sup>10</sup>

$$(18) \quad \int_G z \cdot \frac{2 \sin \frac{\alpha t}{2}}{t} e^{it(u-w)} dX dt = \int_0^{t_1} \frac{2 \sin \frac{\alpha t}{2}}{t} dt \int_E e^{it(u-w)} \cdot z \cdot dX.$$

Now for any positive  $\epsilon$  there exists a  $t_1 > 0$  such that

$$(19) \quad \left| \int_{t_1}^{t_2} \frac{2 \sin \frac{\alpha t}{2}}{t} dt \int_E z \cdot e^{it(u-w)} dX \right| < \frac{\epsilon}{2}$$

\* For the sake of convenience we shall understand a single integral sign to represent a multiple integral where necessary.

\*\* The proof here given is modeled after a similar one of E. L. Dodd (See Annals of Math. 2nd S. Vol. 27, pp. 12-20).

for every  $t_2 > t_1$ , since  $\left| \int_E z \cdot e^{\frac{it(u-\omega)}{2}} dX \right| \leq \left| \int_E z dX \right| \leq 1$ ,

and  $\int_0^\infty \frac{2 \sin \frac{\alpha t}{2}}{t} dt = \pi$  for  $\alpha > 0$ . Furthermore,

$$(20) \quad \left| \int_0^{t_1} \frac{2 \sin \frac{\alpha t}{2}}{t} e^{\frac{it(u-\omega)}{2}} dt \right| \leq \left| \int_0^{t_1} \frac{2 \sin \frac{\alpha t}{2}}{t} dt \right| < 4,$$

and since  $\int_R z dX = 1$ , we can find a rectangular region  $E_1$ ,

such that if  $E$  encloses  $E_1$  and  $E_2$  is that portion of  $E$  not in  $E_1$ ,

$$(21) \quad \left| \int_{E_2} z \cdot dX \right| < \frac{\epsilon}{8} \quad \text{Thus}$$

$$(22) \quad \left| \int_0^{t_1} \frac{2 \sin \frac{\alpha t}{2}}{t} dt \int_{E_2} z \cdot e^{\frac{it(u-\omega)}{2}} dX \right| = \left| \int_{E_2} z dX \int_0^{t_1} \frac{2 \sin \frac{\alpha t}{2}}{t} e^{\frac{it(u-\omega)}{2}} dt \right| < \frac{\epsilon}{2}.$$

Hence, since  $t_2$  and  $E$  may now increase without limit (19) and (22) show the convergence of the  $(n+1)$  fold integral of

$$z \cdot \frac{2 \sin \frac{\alpha t}{2}}{t} \cdot e^{\frac{it(u-\omega)}{2}}.$$

But since  $\left| e^{\frac{it(u-\omega)}{2}} \right| = 1$ ,  $\int_R e^{\frac{it(u-\omega)}{2}} z dX$

exists for all values of  $t$ . Therefore,

$$\int_0^\infty \frac{2 \sin \frac{\alpha t}{2}}{t} dt \int_R z \cdot e^{\frac{it(u-\omega)}{2}} dX$$

exists being equal to the corresponding multiple integral whose existence has just been proved. We have thus established (17) by using the theorem that if the multiple integral and a corresponding iterated integral both exist they are equal.

We can show in a similar manner that

$$\int_R \bar{z} d\mathbf{X} \int_0^\infty \frac{2 \sin \frac{\alpha t}{2}}{t} e^{-it(u-\omega)} dt = \int_0^\infty \frac{2 \sin \frac{\alpha t}{2}}{t} dt \int_R e^{-it(u-\omega)} \cdot \bar{z} \cdot d\mathbf{X},$$

so that finally

$$(23) \quad \int_R \bar{z} d\mathbf{X} \int_{-\infty}^\infty \frac{2 \sin \frac{\alpha t}{2}}{t} e^{it(u-\omega)} dt = \int_{-\infty}^\infty \frac{2 \sin \frac{\alpha t}{2}}{t} dt \int_R \bar{z} e^{it(u-\omega)} d\mathbf{X}.$$

Let  $u_1$  and  $u_2$  approach an intermediate value  $v$  as a limit with  $u_2 > u_1$ . Then  $\alpha \rightarrow dv$  and  $\omega \rightarrow v$  and in the limit

$$(24) \quad P(v) dv = \frac{1}{2\pi} \int_{-\infty}^\infty \frac{2 \sin \frac{tdv}{2}}{t} e^{-itv} dt \int_R e^{itu(x_1, x_2, \dots, x_n)} \cdot \bar{z} \cdot d\mathbf{X}.$$

$$P(v) \text{ exists since } \left| \int_R e^{itu} \cdot \bar{z} \cdot d\mathbf{X} \right| \leq 1 \text{ and}$$

$$\left| \frac{1}{2\pi} \int_{-\infty}^\infty \frac{2 \sin \frac{tdv}{2}}{t} e^{-itv} dt \right| \leq \frac{2}{\pi} \int_0^\infty \frac{\sin \frac{tdv}{2}}{t} dt = 1.$$

Therefore, to within infinitesimals of a higher order, the d.l. of

$$u(x_1, \dots, x_n) \text{ is given by } P(v) dv = \frac{dv}{2\pi} \int_{-\infty}^\infty e^{-itv} dt \int_R e^{itu(x_1, x_2, \dots, x_n)} \cdot \bar{z} d\mathbf{X}$$

$$\text{or } (25) \quad P(v) = \frac{1}{2\pi} \int_{-\infty}^\infty e^{-itv} \varphi(t) dt \text{ where } \varphi(t) = \int_R e^{itu(x_1, x_2, \dots, x_n)} \cdot \bar{z} d\mathbf{X}.$$

An application of Fourier's Integral Theorem<sup>38</sup> to (25) yields finally

$$(26) \quad \varphi(t) = \int_{-\infty}^\infty e^{itv} P(v) dv = \int_R e^{it u(x_1, x_2, \dots, x_n)} \cdot \bar{z} d\mathbf{X},$$

where  $P(v) \equiv 0$  outside the range of applicability.



From (26) we see that  $\mathcal{G}(t)$  is the characteristic function of the d.l. of  $u(x_1, x_2, \dots, x_n)$ .

We now state

\*THEOREM I. If  $u = u(x_1, x_2, \dots, x_n)$  is any function which may have at most a denumerable infinity of discontinuities, of the variables  $x_1, x_2, \dots, x_n$  where the distribution law of  $x_1, x_2, \dots, x_n$  is given by  $f(x_1, x_2, \dots, x_n)$  which is on a certain  $n$ -dimensional manifold  $R$  a single valued, non-negative continuous function

such that  $\int_R f(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n = 1$  then the

characteristic function of the distribution law of  $u$  is given by

$$\mathcal{G}(t) = \int_R e^{it u(x_1, \dots, x_n)} f(x_1, \dots, x_n) dx_1 \dots dx_n.$$

\*\*THEOREM II. Under the conditions of Theorem I, the distribution law of  $u$  is given by

$$P(u) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-it u} \mathcal{G}(t) dt \quad \text{where}$$

$$\mathcal{G}(t) = \int_R e^{it u(x_1, x_2, \dots, x_n)} f(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n.$$

IV. Theorems Regarding Several Functions  $u_j(x_1, x_2, \dots, x_n)$ ,  $j = 1, 2, \dots, r$  : The procedure in the case where we consider several functions  $u_j(x_1, \dots, x_n)$ ,  $j = 1, 2, \dots, r$  of the variables

\*Charlier<sup>3</sup> (Arkiv. Vol. 8) considers a function  $u(x_1, x_2, \dots, x_n)$  which may not be infinite for real  $x_j$ ; nor may the maxima and minima of  $u$  be infinitely dense for any values of the variables.

Kameda<sup>15</sup> (Proc. Vol. 9) considers a function  $u(x_1, x_2, \dots, x_n)$  such that (1)  $u$  must be a continuous function of at least one argument, say  $x_n$ , (2) the derivative of  $u$  with respect to  $x_n$  exists, (3) there exists no interval of  $x_n$  for which  $\frac{\partial u}{\partial x_n}$  is identically zero, (4) the function  $u$  and its derivatives have the same sign in the neighborhood of  $\pm\infty$ .

\*\*Dodd<sup>6</sup> (Annals Vol. 27) considers the distribution of a continuous function  $u(x_1, x_2, \dots, x_n)$ .

$x_1, x_2, \dots, x_n$  is similar to that above.

The probability that  $u_j(x_1, x_2, \dots, x_n)$ ,  $j = 1, 2, \dots, n$ , where the  $u_j$ ,  $j = 1, 2, \dots, n$  and  $x_k$ ,  $k = 1, 2, \dots, n$ , are defined as for the case of a single function  $u$ , satisfy the conditions

$$(27) \quad \begin{cases} u'_1 < u_1 < u''_1 \\ u'_2 < u_2 < u''_2 \\ \dots\dots\dots \\ u'_n < u_n < u''_n \end{cases}$$

is given by

$$(28) \quad \int_B f(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n$$

where the region  $B$  is defined by the set of inequalities (27).

We can avoid the difficulty of integrating over the region  $B$  by introducing the discontinuity factor<sup>88</sup>

$$(29) \quad F = \frac{1}{(2\pi)^n} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \int_{u'_1}^{u''_1} \dots \int_{u'_n}^{u''_n} e^{it_1(\theta_1 - u_1) + it_2(\theta_2 - u_2) + \dots + it_n(\theta_n - u_n)} d\theta_1 \dots d\theta_n dt_1 \dots dt_n$$

$$\text{where } F = 1 \quad \text{for} \quad \begin{cases} u'_1 < u_1 < u''_1 \\ \dots\dots\dots \\ u'_n < u_n < u''_n \end{cases}$$

$$\text{and } F = 0 \quad \text{for} \quad \begin{cases} u'_1 \geq u_1; u_1 \geq u''_1 \\ \dots\dots\dots \\ u'_n \geq u_n; u_n \geq u''_n \end{cases}.$$

We can now say that the probability that  $u_1, u_2, \dots, u_n$  satisfy the conditions (27) is given by

$$(30) \quad \frac{1}{(2\pi)^n} \int_R z dX \int_{-\infty}^{\infty} \int_{u_j}^{u'_j} e^{it_1(\theta_1 - u_1) + it_2(\theta_2 - u_2) + \dots + it_n(\theta_n - u_n)} d\theta_1 \dots d\theta_n dt_1 \dots dt_n.$$

In a manner entirely analogous to the case of a single function

$u$ , we find that

$$(31) \quad P(u_1, u_2, \dots, u_n) = \frac{1}{(2\pi)^n} \int_{-\infty}^{\infty} e^{-it_1 u_1 - it_2 u_2 - \dots - it_n u_n} \varphi(t_1, t_2, \dots, t_n) dt_1 \dots dt_n$$

where

$$(32) \quad \varphi(t_1, t_2, \dots, t_n) = \int_R z \cdot e^{it_1 u_1(x_1, \dots, x_n) + \dots + it_n u_n(x_1, \dots, x_n)} dX.$$

An application of Fourier's Integral Theorem<sup>38</sup> to (31) yields

$$\varphi(t_1, t_2, \dots, t_n) = \int_{-\infty}^{\infty} e^{it_1 u_1 + it_2 u_2 + \dots + it_n u_n} P(u_1, \dots, u_n) du_1 \dots du_n$$

where  $P(u_1, \dots, u_n) \equiv 0$  outside the region of applicability, which shows that  $\varphi(t_1, t_2, \dots, t_n)$  also given as in (32) is the characteristic function of the d.l. of  $u_1, u_2, \dots, u_n$ .

We now state

**THEOREM III.** If  $u_j = u_j(x_1, x_2, \dots, x_n)$ ,  $j = 1, 2, \dots, n$ , which may have a denumerable infinity of discontinuities, are functions of the variables  $x_1, x_2, \dots, x_n$  whose distribution law is given by  $f(x_1, x_2, \dots, x_n)$  which is on a certain  $n$ -dimensional manifold  $R$  a single valued, non-negative continuous function such that  $\int_R f(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n = 1$ , then the characteristic function of the distribution law of  $u_1, u_2, \dots, u_n$  is given by

$$\varphi(t_1, t_2, \dots, t_n) = \int_R e^{it_1 u_1(x_1, \dots, x_n) + \dots + it_n u_n(x_1, \dots, x_n)} \cdot f(x_1, \dots, x_n) dx_1 \dots dx_n.$$

**THEOREM IV.** Under the conditions of Theorem III, the distribution law of  $u_1, u_2, \dots, u_n$  is given by

$$P(u_1, u_2, \dots, u_n) = \frac{1}{(2\pi)^n} \int_{-\infty}^{\infty} e^{-it_1 u_1 - it_2 u_2 - \dots - it_n u_n} \varphi(t_1, \dots, t_n) dt_1 \dots dt_n$$

where

$$\varphi(t_1, t_2, \dots, t_n) = \int_R e^{it_1 u_1(x_1, \dots, x_n) + \dots + it_n u_n(x_1, \dots, x_n)} \cdot f(x_1, \dots, x_n) dx_1 \dots dx_n.$$

## PART 2

*Various Special Cases of the Distribution Problem*

V. *Distribution of the arithmetic mean:*<sup>31</sup> If we take  $u(x_1, \dots, x_n) = x_1 + x_2 + \dots + x_n$  and assume that  $x_1, x_2, \dots, x_n$  are independently distributed each according to the same distribution law, then we find for the distribution of totals

$$(33) \quad P(u) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itu} dt \left( \int_a^b e^{itx} f(x) dx \right)^n, \quad a \leq x \leq b.$$

The substitution  $u = n\bar{x}$  will then yield the distribution of the arithmetic mean.

This result has been derived previously by Poisson,<sup>28</sup> F. Hausdorff<sup>11</sup> and J. O. Irwin.<sup>13</sup>

Hausdorff applied it in particular to find the distribution of means of samples obeying the law  $f(x) = 1/2$  for  $-1 \leq x \leq 1$

and  $f(x) = 0$  elsewhere (a rectangular universe); also to the law  $f(x) = \frac{e^{-|x|}}{2}$ ,  $-\infty \leq x \leq \infty$ . Irwin has applied it to the normal law, Pearson Type III distribution, Pearson Type II distribution and a rectangular universe.

VI. *Distribution of the geometric mean:*<sup>7</sup>

Let  $u = \log x_1 + \log x_2 + \dots + \log x_n$  where  $x_j$ ,  $j=1, 2, \dots, n$  are distributed independently each according to the same distribution law, then

$$(34) \quad P(u) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itu} dt \left( \int_a^b x^{it} f(x) dx \right)^n, \quad 0 \leq a \leq x \leq b.$$

The distribution for the geometric mean  $g$  is obtained from that of  $u$  by the transformation  $u = \log g^n$ .

a. Consider, for example, the case for  $f(x) = \frac{1}{a}$ ,  $0 \leq x \leq a$ .

Then

$$(35) \quad P(u) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{it(n \log a - u)}}{(1+it)^n} dt,$$

where  $n \log a - u \geq 0$ .

From (35) we have

$$(36) \quad P(u) = \frac{(n \log a - u)^{n-1} e^{-(n \log a - u)}}{\Gamma n}$$

since\*  $\int_{-\infty}^{\infty} \frac{e^{ibx} dx}{(1+ix)^n} = \frac{2\pi}{\Gamma n} b^{n-1} e^{-b}$  for  $b > 0$ .

From (36) we obtain

$$(37) \quad D(g) dg = \frac{n g^{n-1}}{a^n \Gamma n} \left( \log \frac{a}{g} \right)^{n-1} dg, \quad 0 \leq g \leq a.$$

The result for  $n=2, 3$  has been given by A. T. Craig.<sup>7</sup>

b. Suppose now that  $f(x) = \frac{x^{p-1} e^{-x}}{\Gamma p}$ ,  $0 \leq x < \infty$ .

Then

$$(38) \quad P(u) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itu} \left( \frac{\Gamma_{p+it}}{\Gamma p} \right)^n dt$$

Let  $p+it = -z$ , then

$$(39) \quad P(u) = \frac{e^{pu}}{(\Gamma p)^n 2\pi i} \int_{-p-i\infty}^{-p+i\infty} e^{uz} (\Gamma z)^n dz.$$

By a method similar to that used for the case of the general-

\* MacRobert,<sup>20</sup> p. 67.

ized variance (see Section XIII), we may show that

$$\lim_{z \rightarrow \infty} |z^m e^{uz} (\sqrt{z})^n| \rightarrow 0,$$

so that the integral converges and

$$(40) \quad P(u) = -\frac{e^{pu}}{(\sqrt{p})^n 2\pi i} \int_C e^{zu} (\sqrt{z})^n dz,$$

where  $C$  is the contour bounded by the line  $x = -p$  and that part of the circle  $|z| = m + \frac{1}{2}$ ,  $m \rightarrow \infty$  which lies to the right of the straight line. The contour is traversed in a counter-clockwise direction.

$$\text{Now } (\sqrt{z})^n = \frac{(-1)^n \pi^n}{\sin \pi z (\sqrt{z+1})^n} \text{ so that we may also write}$$

$$(41) \quad P(u) = -\frac{e^{pu}}{(\sqrt{p})^n 2\pi i} \int_C \frac{(-1)^n \pi^n e^{uz}}{\sin \pi z (\sqrt{z+1})^n} dz.$$

The poles of the integrand are of the  $n^{\text{th}}$  order and are those of  $(\sqrt{z})^n$  viz.,  $z = \lambda$ ,  $\lambda = 0, 1, 2, \dots$ . Since the contour is traversed in a counter-clockwise manner, the value of the integral is  $2\pi i$  times the sum of the residues at the poles within the contour so that

$$(42) \quad P(u) = \frac{e^{pu}}{(\sqrt{p})^n} \sum_{\lambda=0}^{\infty} \frac{(-1)^{n+n\lambda}}{(n-1)!} \left[ \frac{d^{n-1}}{dz^{n-1}} \frac{e^{uz}}{(\sqrt{z+1})^n} \right]_{z=\lambda}$$

or

$$(43) \quad D(g) = \frac{n g^{np-1}}{\sqrt{n} (\sqrt{p})^n} \sum_{\lambda=0}^{\infty} (-1)^{n+n\lambda} \left[ \frac{d^{n-1}}{dz^{n-1}} \frac{g^{nz}}{(\sqrt{z+1})^n} \right]_{z=\lambda}$$

c. If instead of assuming the  $x_j$ 's each satisfy the same distribution law, we assume  $x_j$  to be distributed according to

$$f(x_j) = \frac{x_j^{p_j-1} e^{-x_j}}{\Gamma p_j} \quad \text{where none of the } p_j \text{'s are equal}$$

or differ by an integer, then

$$(44) \quad P(u) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-it u} \frac{\prod_{j=1}^n \sqrt{p_j + it}}{\prod_{j=1}^n \sqrt{p_j}} dt,$$

or

$$(45) \quad P(u) = \frac{1}{\prod_{j=1}^n \sqrt{p_j} \cdot 2\pi i} \int_{-\infty i}^{\infty i} e^{u z} \frac{\prod_{j=1}^n \sqrt{p_j - z}}{\prod_{j=1}^n \sqrt{p_j - z}} dz.$$

The same results as to the convergence of the integral and the contour may be shown with respect to this integrand as for Section VI b.

The value of

$$J = \int_C e^{u z} \frac{\prod_{j=1}^n \sqrt{p_j - z}}{\prod_{j=1}^n \sqrt{p_j - z}} dz$$

is  $2\pi i$  times the sum of the residues within the contour bounded by the  $y$ -axis and that part of the circle  $|z| = m + \frac{1}{2}$ ,  $m \rightarrow \infty$ , which lies to the right of this line.

For the pole  $z = p_j + \kappa$ ,  $\kappa = 0, 1, 2, \dots$  the residue is

$$(-1)^\kappa \frac{e^{u(p_j + \kappa)}}{\kappa!} \sqrt{p_1 - p_j - \kappa} \sqrt{p_2 - p_j - \kappa} \cdots \sqrt{p_{j-1} - p_j - \kappa} \sqrt{p_{j+1} - p_j - \kappa} \cdots \sqrt{p_n - p_j - \kappa}$$

therefore,

$$(46) \quad P(u) = \frac{1}{\prod_{j=1}^n \sqrt{p_j}} \left\{ \sum_{j=1}^n \sum_{\kappa=0}^{\infty} \frac{(-1)^\kappa e^{u(p_j + \kappa)}}{\kappa!} \prod_{\substack{k=1 \\ k \neq j}}^n \sqrt{p_k - p_j - \kappa} \right\}$$

where  $\prod_{\substack{k=1 \\ k \neq j}}^n$  means that in the product  $\kappa$  takes all the values  $1, 2, \dots, n$  except  $j$ .

Finally,

$$(47) \quad D(g) = \frac{1}{\prod_{j=1}^n \sqrt{p_j}} \left\{ \sum_{j=1}^n \sum_{\kappa=0}^{\infty} \frac{(-1)^\kappa g^{p_j + \kappa}}{\kappa!} \prod_{\substack{k=1 \\ k \neq j}}^n \sqrt{p_k - p_j - \kappa} \right\}.$$

d. Suppose that in the previous case  $p_j = p + \frac{j-1}{n}$ ,  $j = 1, 2, \dots, n$ .

Since  $\sqrt{p} \sqrt{p + \frac{1}{n}} \cdots \sqrt{p + \frac{n-1}{n}} = n^{\frac{1}{2} - np} (2\pi)^{\frac{n-1}{2}} \sqrt{np}$   
for this case

$$(48) \quad P(u) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-itu} n^{\frac{1}{2} - n(p+it)} (2\pi)^{\frac{n-1}{2}} \sqrt{np+nit}}{n^{\frac{1}{2} - np} (2\pi)^{\frac{n-1}{2}} \sqrt{np}} dt.$$

Let  $np + nit = -z$ , then

$$(49) \quad P(u) = \frac{e^{up} n^{np}}{n \sqrt{np} \cdot 2\pi i} \int_{-np-i\infty}^{-np+i\infty} (e^{\frac{u}{n} \cdot n})^z \sqrt{-z} dz.$$

Now it may be shown that

$$\frac{1}{2\pi i} \int_{-a-i\infty}^{-a+i\infty} u^z \sqrt{-z} dz = e^{-u}$$

where  $a > 0$  and  $-\frac{\pi}{2} < \arg u < \frac{\pi}{2}$  (See MacRobert,<sup>20</sup> p. 151.)

Therefore

$$(50) \quad P(u) = \frac{e^{up} n^{np}}{n \sqrt{np}} e^{-\frac{u}{n}}$$

Substituting  $g^n = e^u$  we obtain for the distribution of  $g$

$$(51) \quad D(g) = \frac{n^{np} g^{np-1} e^{-ng}}{\sqrt{np}}, \quad 0 \leq g \leq \infty.$$

In other words, the distribution of the geometric mean of  $n$  independent variables respectively satisfying the distribution law

$$\frac{x^{p-1} e^{-x}}{\sqrt{p}}; \frac{x^{p+\frac{1}{n}-1} e^{-x}}{\sqrt{p+\frac{1}{n}}}; \dots; \frac{x^{p+\frac{n-1}{n}-1} e^{-x}}{\sqrt{p+\frac{n-1}{n}}}, \quad 0 \leq x \leq \infty$$

is the same as the distribution of the arithmetic mean of  $n$  independent variables each satisfying the Pearson Type III distribu-



tion law

$$f(x) = \frac{x^{p-1} e^{-x}}{\Gamma(p)}, \quad 0 \leq x < \infty.$$

e. For the case where  $p_j = \frac{N-j}{2}$ ,  $j = 1, 2, \dots, n$ , see the discussion for the generalized variance (see Section XIII).

VII. *Lemma*: The following geometrical considerations will for certain cases simplify the problem of finding the distribution of statistical parameters calculated about a sample mean.

Consider the sample as a point or points (for multi-variate distributions) in an  $n$ -dimensional Euclidean space. (This method has been employed to great advantage by R. A. Fisher<sup>8</sup> and others.) Then, if the probability density at any point (the probability for that particular combination of values to occur) is a function of the distance from the origin, the mean value of a function of the distance from the origin and of other geometric invariants of the system for  $x_j, y_j, \dots, j = 1, 2, \dots, n$  satisfying the conditions  $\sum_{j=1}^n x_j = 0, \sum_{j=1}^n y_j = 0, \dots$  will be the same as for the same function for independent variables in  $n-1$  dimensional space. Since the important element is the distance from the origin and the integration is to be carried out over an  $n-1$  dimensional space, the final result is independent of the fact that the whole system is immersed in an  $n$ -dimensional space.

As an illustration, let us consider the following distributions which have been derived by various methods.

VIII. *Distribution of variance of a sample of  $n$  from a normal population*.<sup>8, 28, 34, 36</sup>

Let  $u = x_1^2 + x_2^2 + \dots + x_{n-1}^2$  where the  $x_j$  are distributed according to  $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}, -\infty \leq x < \infty$ .

Then

$$(52) \quad \varphi(t) = \left[ \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2} + itx} dx \right]^{n-1} = \frac{1}{(1 - 2\sigma^2 t^2)^{\frac{n-1}{2}}}.$$

(Compare Rider,<sup>31</sup> *Annals* p. 600; Romanovsky,<sup>34</sup> *Metron* p. 6.) Therefore, the distribution of

$$V = x_1^2 + x_2^2 + \dots + x_n^2 = nS^2, \quad \text{where } \sum_{j=1}^n x_j = 0,$$

is given by

$$(53) \quad P(v) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-itv} dt}{(1 - 2\sigma^2 it)^{\frac{n-1}{2}}} = \frac{v^{\frac{n-3}{2}} e^{-\frac{v}{2\sigma^2}}}{(2\sigma^2)^{\frac{n-1}{2}} \sqrt{\frac{n-1}{2}}}$$

(see MacRobert,<sup>20</sup> p. 67.)

We thus have

$$(54) \quad D(s^2) ds^2 = \left(\frac{n}{2\sigma^2}\right)^{\frac{n-1}{2}} \frac{(s^2)^{\frac{n-3}{2}} e^{-\frac{ns^2}{2\sigma^2}} ds^2}{\sqrt{\frac{n-1}{2}}} \text{ as is well known.}$$

IX. *Distribution of the  $\chi^2$  of Goodness of Fit Test:*<sup>22, 24</sup> Consider

$$\chi^2 = \sum_{j,k=1}^n \frac{R_{jk}}{R} \frac{x_j x_k}{\sigma_j \sigma_k},$$

where  $R = |\rho_{jk}|$ ,  $\rho_{jj} = 1$  and  $R_{jk}$  is the cofactor of  $\rho_{jk}$  in  $R$  so that  $|R_{jk}| = R^{n-1}$  and  $x_1, x_2, \dots, x_n$  are distributed according to

$$\frac{e^{-\frac{\chi^2}{2}}}{(2\pi)^{n/2} \sigma_1 \sigma_2 \dots \sigma_n R^{1/2}}$$

Therefore

$$(55) \quad P(\chi^2) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-it\chi^2} dt \int \frac{e^{-\frac{1}{2R} \sum_{j,k=1}^n \frac{R_{jk}(1-2it)x_j x_k}{\sigma_j \sigma_k}} dx_1 dx_2 \dots dx_n}{(2\pi)^{n/2} \sigma_1 \sigma_2 \dots \sigma_n R^{1/2}} \\ = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-it\chi^2} dt \int \frac{e^{-\frac{1}{2} \sum_{j,k=1}^n R_{jk}(1-2it)x_j x_k}}{(2\pi)^{n/2} R^{-\frac{n-1}{2}}} dx_1 \dots dx_n.$$

$$\begin{aligned} \therefore P(\chi^2) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-it\chi^2} \frac{|R_{jK}|^{1/2}}{|R_{jK}(1-2it)|^{1/2}} dt \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-it\chi^2} dt}{(1-2it)^{n/2}} \end{aligned}$$

and we have finally,

$$(56) \quad P(\chi^2) d\chi^2 = \frac{1}{\sqrt{\frac{n}{2}}} \left(\frac{\chi^2}{2}\right)^{\frac{n-2}{2}} e^{-\frac{\chi^2}{2}} d\left(\frac{\chi^2}{2}\right).$$

If we restrict the  $x_j$  in  $\chi^2$  to satisfy  $\sum_{j=1}^n x_j = 0$ , then

from the preceding, it is clear that  $\varphi(t) = \frac{1}{(1-2it)^{n/2}}$  and now

$$(57) \quad P(\chi^2) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-it\chi^2} dt}{(1-2it)^{n/2}}$$

or

$$(58) \quad P(\chi^2) d\chi^2 = \frac{1}{\sqrt{\frac{n-1}{2}}} \left(\frac{\chi^2}{2}\right)^{\frac{n-3}{2}} e^{-\frac{\chi^2}{2}} d\left(\frac{\chi^2}{2}\right).$$

This latter case is the one commonly met with in actual practice and is equivalent to the case wherein the expected values are adjusted according to the total in the sample.

X. *Simultaneous distribution of variances and correlation coefficient of a sample of  $n$  from a bi-variate normal population.*<sup>8</sup>

This is a special case of the problem of finding the simultaneous distribution of the variances and covariances from an  $n$ -variate normal population which has been solved by J. Wishart.<sup>42</sup> The same method is applicable to the general case, but for its own interest and for the sake of simplicity this special case will be considered.

$$\text{Let } u_1 = \frac{\sum_{j=1}^n x_j^2}{2(1-\rho^2)\sigma_x^2}; \quad u_2 = \frac{\rho \sum_{j=1}^n x_j y_j}{(1-\rho^2)\sigma_x \sigma_y}; \quad u_3 = \frac{\sum_{j=1}^n y_j^2}{2(1-\rho^2)\sigma_y^2}$$

where  $x_j$  and  $y_j$  are distributed according to

$$\frac{1}{2\pi \sigma_x \sigma_y \sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)} \left[ \frac{x_j^2}{\sigma_x^2} - 2\rho \frac{x_j y_j}{\sigma_x \sigma_y} + \frac{y_j^2}{\sigma_y^2} \right]}$$

Now consider

$$\begin{aligned} (59) \quad J &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2(1-\rho^2)} \left[ \frac{(1-it_1)x^2}{\sigma_x^2} - \frac{2\rho(1+it_2)xy}{\sigma_x \sigma_y} + \frac{(1-it_3)y^2}{\sigma_y^2} \right]}}{2\pi \sigma_x \sigma_y \sqrt{1-\rho^2}} dx dy \\ &= \frac{(1-\rho^2)^{1/2}}{\left[ (1-it_1)(1-it_3) - \rho^2(1+it_2)^2 \right]^{1/2}} \\ &= \frac{(1-\rho^2)^{1/2}}{\begin{vmatrix} 1-it_1 & \rho(1+it_2) \\ \rho(1+it_2) & 1-it_3 \end{vmatrix}}^{1/2} \end{aligned}$$

Therefore, if we add the conditions  $\sum_{j=1}^n x_j = 0$ ;  $\sum_{j=1}^n y_j = 0$ , in which case

$$u_1 = \frac{n S_x^2}{2(1-\rho^2)\sigma_x^2}; \quad u_2 = \frac{n \rho S_x S_y}{(1-\rho^2)\sigma_x \sigma_y}; \quad u_3 = \frac{n S_y^2}{2(1-\rho^2)\sigma_y^2} \quad \text{and}$$

$$(60) \quad g(t_1, t_2, t_3) = \frac{(1-\rho^2)^{\frac{n-1}{2}}}{\left[ (1-it_1)(1-it_3) - \rho^2(1+it_2)^2 \right]^{\frac{n-1}{2}}}.$$

Therefore

$$(61) \quad P(u_1, u_2, u_3) = \frac{(1-\rho^2)^{\frac{n-1}{2}}}{(2\pi)^3} \iiint_{-\infty}^{+\infty} \frac{e^{-it_1 u_1 - it_2 u_2 - it_3 u_3}}{\left[ (1-it_1)(1-it_3) - \rho^2(1+it_2)^2 \right]^{\frac{n-1}{2}}} dt_1 dt_2 dt_3.$$

Integrating with respect to  $t_1$ , we find

$$(62) \int_{-\infty}^{\infty} \frac{e^{-it_1 u_1} dt_1}{[(1-it_1)(1-it_3)-\rho^2(1+it_3)^2]^{\frac{n-1}{2}}} = \frac{2\pi}{\sqrt{n-2}} \frac{u_1^{\frac{n-3}{2}} e^{-u_1 \left[1 - \frac{\rho^2(1+it_3)^2}{1-it_3}\right]}}{(1-it_3)^{\frac{n-1}{2}}}.$$

Integrating with respect to  $t_2$ , we find

$$(63) \int_{-\infty}^{\infty} e^{-it_2 \left(u_2 - \frac{u_1 \rho^2}{1-it_3} - it_2 \left(u_2 - \frac{2u_1 \rho^2}{1-it_3}\right)\right)} dt_2 = \sqrt{\frac{\pi(1-it_3)}{u_1 \rho^2}} e^{-\left(u_2 - \frac{2u_1 \rho^2}{1-it_3}\right) \frac{1-it_3}{4u_1 \rho^2}}$$

Integrating with respect to  $t_3$ , we find

$$(64) \int_{-\infty}^{\infty} \frac{e^{-it_3 \left(u_3 - \frac{u_2^2}{4u_1 \rho^2}\right)}}{(1-it_3)^{\frac{n-2}{2}}} dt_3 = \frac{2\pi}{\sqrt{n-2}} \left(u_3 - \frac{u_2^2}{4u_1 \rho^2}\right)^{\frac{n-4}{2}} e^{-\left(u_3 - \frac{u_2^2}{4u_1 \rho^2}\right)}$$

using the facts that 
$$\int_{-\infty}^{\infty} e^{-ax^2 \pm 2bx} dx = \sqrt{\frac{\pi}{a}} e^{b^2/a}$$

and 
$$\int_{-\infty}^{\infty} \frac{e^{-ibx} dx}{(1-ix)^n} = \frac{2\pi}{\sqrt{n}} b^{n-1} e^{-b}.$$

Therefore we finally find that

$$(65) P(u_1, u_2, u_3) = \frac{(1-\rho^2)^{\frac{n-1}{2}}}{2\rho\sqrt{\pi}\sqrt{\frac{n-1}{2}}\sqrt{\frac{n-2}{2}}} e^{-\left(u_1 - u_2 + u_3\right)} \left(1 - \frac{u_2^2}{4u_1 u_3 \rho^2}\right)^{\frac{n-4}{2}} u_1^{\frac{n-4}{2}} u_3^{\frac{n-4}{2}},$$

or

$$(66) D(s_x, r, s_y) ds_x dr ds_y = \frac{n(1-\rho^2)^{\frac{n-1}{2}} s_x^{\frac{n-2}{2}} s_y^{\frac{n-2}{2}} e^{-\frac{n}{2(1-\rho^2)} \left[\frac{s_x^2}{\sigma_x^2} - \frac{2\rho r s_x s_y}{\sigma_x \sigma_y} + \frac{s_y^2}{\sigma_y^2}\right]}}{(1-\rho^2)^{\frac{n-1}{2}} \pi \sqrt{n-2} \sigma_x^{n-1} \sigma_y^{n-1}}$$

since 
$$2^{\frac{n-3}{2}} \sqrt{\frac{n-1}{2}} \sqrt{\frac{n-2}{2}} = \pi^{\frac{1}{2}} \sqrt{n-2}$$

and 
$$\frac{\partial(u_1, u_2, u_3)}{\partial(s_x, r, s_y)} = \frac{\pi^3 \rho^3 s_x^2 s_y^2}{(1-\rho^2)^3 \sigma_x^3 \sigma_y^3}.$$

X. The distribution of the covariance of a sample of  $n$  from a bi-variate normal population:<sup>25</sup>

$$\text{Let } u = \frac{\rho}{(1-\rho^2) \sigma_x \sigma_y} \sum_{j=1}^n x_j y_j$$

where  $x_j$  and  $y_j$  are distributed according to

$$\frac{1}{2\pi \sigma_x \sigma_y \sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)} \left[ \frac{x_j^2}{\sigma_x^2} - 2 \frac{\rho x_j y_j}{\sigma_x \sigma_y} + \frac{y_j^2}{\sigma_y^2} \right]}$$

Consider

$$\begin{aligned} (67) \quad J &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2(1-\rho^2)} \left[ \frac{x^2}{\sigma_x^2} - 2 \frac{(1+it)\rho xy}{\sigma_x \sigma_y} + \frac{y^2}{\sigma_y^2} \right]}}{2\pi \sigma_x \sigma_y \sqrt{1-\rho^2}} dx dy \\ &= \frac{(1-\rho^2)^{1/2}}{[1-\rho^2(1+it)^2]^{1/2}}. \end{aligned}$$

If we impose the conditions

$$\sum_{j=1}^n x_j = 0; \quad \sum_{j=1}^n y_j = 0 \quad \text{so that } u = \frac{n\rho S_{xy}}{(1-\rho^2) \sigma_x \sigma_y}$$

then

$$(68) \quad g(t) = \frac{(1-\rho^2)^{\frac{n-1}{2}}}{[1-\rho^2(1+it)^2]^{\frac{n-1}{2}}}$$

and

$$(69) \quad P(u) = \frac{(1-\rho^2)^{\frac{n-1}{2}}}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-itu} dt}{[1-\rho^2(1+it)^2]^{\frac{n-1}{2}}}$$

Consider

$$(70) \quad I = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-itu} dt}{\{[1-\rho(1+it)][1+\rho(1+it)]\}^{\frac{n-1}{2}}}$$

Let  $1 - \rho(1 + it) = -\frac{\rho z}{u}$  so that

$$(71) \quad I = \frac{u^{\frac{n-3}{2}} e^{\frac{u(\rho-1)}{\rho}}}{(2\rho)^{\frac{n-1}{2}} 2\pi i} \int_{-\frac{1-\rho}{\rho}u+i\infty}^{-\frac{1-\rho}{\rho}u-i\infty} \frac{e^{-z} dz}{(-z)^{\frac{n-1}{2}} \left(1 + \frac{z\rho}{2u}\right)^{\frac{n-1}{2}}}$$

Since we may show that

$$\lim_{z \rightarrow \infty} \left| z^m \frac{e^{-z}}{(-z)^{\frac{n-1}{2}} \left(1 + \frac{z\rho}{2u}\right)^{\frac{n-1}{2}}} \right| \rightarrow 0$$

the integral is convergent and we may write

$$(72) \quad I = -\frac{u^{\frac{n-3}{2}} e^{\frac{u\rho-1}{\rho}}}{(2\rho)^{\frac{n-1}{2}} 2\pi i} \int_{\infty}^{(0+)} \frac{e^{-z} dz}{(-z)^{\frac{n-1}{2}} \left(1 + \frac{z\rho}{2u}\right)^{\frac{n-1}{2}}},$$

where  $\int_{\infty}^{(0+)}$  means that the path of integration starts at infinity on the real axis, encircles the origin in the positive direction and returns to the starting point. (See Whittaker and Watson,<sup>40</sup> pp. 239, 333.)

Since  $\frac{1-\rho}{\rho} < \frac{z}{\rho}$ , the point  $z = -\frac{z}{u\rho}$  is outside the contour so that

$$(73) \quad -\frac{1}{2\pi i} \int_{\infty}^{(0+)} \frac{e^{-z} dz}{(-z)^{\frac{n-1}{2}} \left(1 + \frac{z\rho}{2u}\right)^{\frac{n-1}{2}}} = \frac{e^{\frac{u}{\rho}} W_{0, -\frac{n-2}{2}}\left(\frac{2u}{\rho}\right)}{\sqrt{\frac{n-1}{2}}},$$

where  $W_{\kappa, m}(z)$  is the confluent hypergeometric function.<sup>40</sup>

Also, since  $W_{0, m}(z) = W_{0, -m}(z)$  we have finally

$$(74) \quad P(u) = \frac{(1-\rho^2)^{\frac{n-1}{2}} u^{\frac{n-3}{2}} e^{\frac{u}{\rho}} W_{0, \frac{n-2}{2}}\left(\frac{2u}{\rho}\right)}{\sqrt{\frac{n-1}{2}} (2\rho)^{\frac{n-1}{2}}}$$

If we start with the following definition for the Bessel Function of the second kind and imaginary argument<sup>26, 27</sup>

$$(75) \quad K_m(x) = \frac{\sqrt{\pi} x^m}{2^m \Gamma(m + \frac{1}{2})} \int_1^\infty e^{-xt} (t^2 - 1)^{m - \frac{1}{2}} dt$$

then it is possible to show that  $K_m(x) = \sqrt{\pi} x^{-\frac{1}{2}} 2^{-\frac{1}{2}} W_{0,m}(2x)$ , so that

$$(76) \quad P(u) = \frac{(1-\rho^2)^{\frac{n-1}{2}} u^{\frac{n-2}{2}} e^{-\frac{u}{\rho}} K_{\frac{n-2}{2}}\left(\frac{u}{\rho}\right)}{\sqrt{\pi} \sqrt{\frac{n-1}{2}} 2^{\frac{n-2}{2}} \rho^{\frac{n}{2}}}.$$

$$\text{If we finally set } v = \frac{u}{\rho} = \frac{n \cdot S_{xy}}{(1-\rho^2) \sigma_x \sigma_y},$$

we find for the distribution of  $v$ ,

$$(77) \quad D(v) dv = \frac{(1-\rho^2)^{\frac{n-1}{2}} e^{-\rho v} v^{\frac{n-2}{2}} K_{\frac{n-2}{2}}(v) dv}{\sqrt{\pi} 2^{\frac{n-2}{2}} \sqrt{\frac{n-1}{2}}},$$

which is the form found by K. Pearson, G. B. Jeffery, F.R. S. and E. M. Elderton.<sup>25</sup>

XII. Do  $N$  samples, each of  $n$ -categories, come from the same  $n$ -variate normal parent?<sup>28</sup> Consider

$$\chi^2 = \sum_{j=1}^N \sum_{j,k=1}^n \frac{R_{jk}}{R} \cdot \frac{x_{j\alpha} x_{k\alpha}}{\sigma_j \sigma_k} \quad \text{where the simul-}$$

aneous distribution of  $x_1, x_2, \dots, x_n$  is given by

$$(78) \quad \frac{e^{-\frac{1}{2R} \sum_{j,k=1}^n R_{jk} \frac{x_j x_k}{\sigma_j \sigma_k}}}{(2\pi)^{n/2} \sigma_1 \sigma_2 \dots \sigma_n R^{1/2}},$$

where  $R_{jk}$  denotes the cofactor corresponding to  $\rho_{jk}$  in the determinant  $R = |\rho_{jk}|$  of the population correlations and  $\sigma_j$  is the standard deviation of the  $j^{\text{th}}$  variate.



Consider

$$J = \int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2R} \sum_{j,k=1}^n (1-2it) \frac{R_{jk} x_j x_k}{\sigma_j \sigma_k}}}{(2\pi)^{n/2} \sigma_1 \sigma_2 \dots \sigma_n R^{1/2}} dx_1 dx_2 \dots dx_n$$

$$= \frac{|R_{jk}|^{1/2}}{|R_{jk}(1-2it)|^{1/2}} = \frac{1}{(1-2it)^{n/2}}$$

If we impose the conditions  $\sum_{j=1}^n x_{j\alpha} = 0$ ,  $\alpha = 1, 2, \dots, N$  and  $\sum_{\alpha=1}^N x_{j\alpha} = 0$ ,  $j = 1, 2, \dots, n$  then from the previous results the characteristic function for the distribution of  $\chi^2$  becomes

$$g(t) = \frac{1}{(1-2it)^{\frac{(n-1)(N-1)}{2}}}$$

and the distribution for  $\chi^2$  is

$$(79) \quad P(\chi^2) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-it\chi^2}}{(1-2it)^{\frac{(n-1)(N-1)}{2}}} dt = \frac{(\chi^2)^{\frac{(n-1)(N-1)-2}{2}} e^{-\frac{\chi^2}{2}}}{2^{\frac{(n-1)(N-1)}{2}} \sqrt{\frac{(n-1)(N-1)}{2}}}$$

This case is equivalent to applying the  $\chi^2$  test to a contingency table. If the table has  $n$  rows and  $c$  columns then the value of  $n'$  to be used in Elderton's tables of "Goodness of Fit" is<sup>9</sup>  $n' = (n-1)(c-1) + 1$  [as we saw in Section IX, equation 58, the distribution for  $\chi^2$  has an exponent  $\frac{n'-3}{2}$  (our  $n$  is equal to the  $n'$  of the table) and the exponent in the distribution above is  $\frac{(n-1)(N-1)-2}{2}$  ].

XIII. Distribution of the generalized variance of a sample of  $N$  from an  $n$ -variate normal population:<sup>41</sup> One of the gen-

eralizations considered by Wilks is that of the sample variance. For a sample of  $N$  from an  $n$  variate normal population the generalized sample variance is defined to be the determinant  $|a_{jk}|$

where  $a_{jk} = a_{kj} = \frac{1}{N} \sum_{\alpha=1}^N (x_{j\alpha} - \bar{x}_j)(x_{k\alpha} - \bar{x}_k)$ ,  $j, k = 1, 2, \dots, n$  and  $\bar{x}_j = \frac{1}{N} \sum_{\alpha=1}^N x_{j\alpha}$ . Wilks has given the distribution of  $u = |a_{jk}|$  as an  $(n-1)$ -tuple integral and has obtained the explicit form of the distribution for  $n = 1, 2$ .

By employing the theory of characteristic functions we are enabled to express the distribution of  $u$  as a single integral and find the explicit form for any value of  $n$ .

The simultaneous distribution of the  $a_{jk}$  defined above is given<sup>42</sup> by

$$(80) \quad \frac{|A_{jk}|^{\frac{N-1}{2}} e^{-\sum_{j,k=1}^n A_{jk} a_{jk}} |a_{jk}|^{\frac{N-n-2}{2}}}{\pi^{\frac{n(n-1)}{4}} \prod_{j=1}^n \sqrt{\frac{N-j}{2}}}$$

where  $|A_{jk}|$  is the  $n$ -th order determinant of elements

$A_{jk} = \frac{N R_{jk}}{2 \sigma_j \sigma_k R}$ , where  $R_{jk}$  is the cofactor of  $\rho_{jk}$  in the determinant of parent correlations  $R = |\rho_{jk}|$ .

If we write  $A_{jk} = N B_{jk}$  and  $a_{jk} = \frac{b_{jk}}{N}$ , the distribution of the  $b_{jk}$ 's is

$$(81) \quad \frac{|B_{jk}|^{\frac{N-1}{2}} e^{-\sum_{j,k=1}^n B_{jk} b_{jk}} |b_{jk}|^{\frac{N-n-2}{2}}}{\pi^{\frac{n(n-1)}{4}} \prod_{j=1}^n \sqrt{\frac{N-j}{2}}}$$

For the sake of concreteness and the better to follow the dis-

cussion for the general case, we shall first consider the cases  $n = 3, 4$  in detail.

Case 1,  $n=3$ : Let  $\xi = \log \ell$  where we write  $\ell = |\ell_{jk}|$ . The distribution of  $\xi$  is then given by

$$(82) \quad P(\xi) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-it\xi} dt \int \frac{|B_{jk}| e^{\frac{N-1}{2} - \sum_{j,k=1}^n B_{jk} \ell_{jk} \frac{N-5+2it}{2}} \ell}{\pi^{3/2} \sqrt{\frac{N-1}{2}} \sqrt{\frac{N-2}{2}} \sqrt{\frac{N-3}{2}}} d\ell_n$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-it\xi} B^{-it} \sqrt{\frac{N-1}{2}+it} \sqrt{\frac{N-2}{2}+it} \sqrt{\frac{N-3}{2}+it} dt}{\sqrt{\frac{N-1}{2}} \sqrt{\frac{N-2}{2}} \sqrt{\frac{N-3}{2}}},$$

where  $B = |B_{jk}|$ . (Compare Wilks,<sup>41</sup> *Biometrika* Vol. 24, p. 477, equation 10.)

Let  $\frac{N-3}{2} + it = -z$ , then

$$(83) \quad P(\xi) = \frac{\xi(B e^{\xi})^{\frac{N-3}{2}}}{\sqrt{\frac{N-1}{2}} \sqrt{\frac{N-2}{2}} \sqrt{\frac{N-3}{2}} 2\pi i} \int_{-\frac{N-3}{2}-i\infty}^{-\frac{N-3}{2}+i\infty} e^{\xi z} B^z \sqrt{1-z} \sqrt{\frac{1}{2}-z} \sqrt{\frac{1}{2}-z} dz.$$

The integral is taken along the line  $x = -\frac{N-3}{2}$  and since  $N > 3$  (since otherwise the distribution of the  $a_{jk}$ 's is nugatory) all the poles of the integrand are to the right of the line  $x = -\frac{N-3}{2}$ .

Now  $\sqrt{1-z} = -z \sqrt{-z}$  so that  $\sqrt{1-z} \sqrt{\frac{1}{2}-z} \sqrt{\frac{1}{2}-z} = -z (\sqrt{-z})^2 \sqrt{\frac{1}{2}-z}$

but  $\sqrt{-z} = -\frac{\pi}{\sin \pi z \sqrt{1+z}}$  so that  $(\sqrt{-z})^2 = \frac{\pi^2}{\sin^2 \pi z \cdot (\sqrt{1+z})^2}$ .

Now

$$\lim_{z \rightarrow \infty} \left| \frac{1}{(\sqrt{z+1})^2} \right| = \lim_{z \rightarrow \infty} \left| \frac{e^{2z - 2z \log z}}{2\pi z^3} \right|$$

If we set  $z = re^{i\theta}$

$$\lim_{z \rightarrow \infty} \left| \frac{1}{(\sqrt{z+1})^2} \right| = \lim_{r \rightarrow \infty} \frac{e^{2r \cos \theta - 2r \cos \theta \log r + 2r \sin \theta}}{2\pi r^3}$$

$$\text{Also } \sqrt{\frac{1}{2} - z} = \frac{\pi}{\cos \pi z \sqrt{z + \frac{1}{2}}}$$

$$\text{and } \lim_{z \rightarrow \infty} \left| \frac{1}{\sqrt{z + \frac{1}{2}}} \right| = \lim_{r \rightarrow \infty} \frac{e^{r \cos \theta - r \cos \theta \log r + \theta r \sin \theta}}{\pi^{1/2} r^{3/2}}$$

$$\text{We also have that } \lim_{z \rightarrow \infty} |\sin \pi z| \leq \lim_{z \rightarrow \infty} e^{\pm \pi r \sin \theta}$$

according as  $\sin \theta$  is positive or negative and that

$$\lim_{z \rightarrow \infty} |\cos \pi z| \leq \lim_{r \rightarrow \infty} e^{\pm \pi r \sin \theta}$$

according as  $\sin \theta$  is positive or negative.

We find therefore that finally,

$$\lim_{z \rightarrow \infty} \left| e^{\frac{1}{2} z} B^z \sqrt{1-z} \sqrt{\frac{1}{2}-z} \sqrt{-z} \right| \leq \sqrt{2\pi} \lim_{r \rightarrow \infty} \frac{e^{r \cos \theta [\frac{1}{2} + \log B + 3 - 3 \log r] + r \sin \theta [3\theta + 3\pi]}}{r^2}$$

according as  $\sin \theta$  is positive or negative.

Therefore if  $\frac{\pi}{2} \geq \theta \geq \epsilon$ ;  $-\frac{\pi}{2} \leq \theta \leq -\epsilon$  ; or if

$$-\frac{N-3}{2} \leq r \cos \theta \leq 0,$$

$z^\pi e^{\frac{1}{2} z} B^z \sqrt{1-z} \sqrt{\frac{1}{2}-z} \sqrt{-z}$  tends uniformly to zero as  $z$  tends

to infinity and the integral is uniformly convergent.\*

Next if  $-\epsilon \leq \theta \leq \epsilon$  let  $z = \rho_m e^{i\theta}$  where  $\rho_m = m + \frac{1}{2}$  and  $m$  is an integer. Then,

$$\lim_{z \rightarrow \infty} \left| e^{\xi z} B^z \sqrt{1-z} \sqrt{\frac{1}{2}-z} \sqrt{-z} \right| \leq M_1 M_2 \lim_{m \rightarrow \infty} \frac{e^{\rho_m \cos \theta [\xi + \log B + 3 - 3 \log \rho_m] + \rho_m \sin \theta |\eta|}}{\rho_m^2}$$

where  $2M_1 \geq |\csc \pi z|$ ;  $2M_2 \geq |\sec \pi z|$ . \*\*

Therefore  $z^n e^{\xi z} B^z \sqrt{1-z} \sqrt{\frac{1}{2}-z} \sqrt{-z}$  tends to zero uniformly as  $m$  tends to infinity.

We can now write

$$(84) \quad P(\xi) = - \frac{(Be^\xi)^{\frac{N-3}{2}}}{\sqrt{\frac{N-1}{2}} \sqrt{\frac{N-2}{2}} \sqrt{\frac{N-3}{2}} 2\pi i} \int_C e^{\xi z} B^z \sqrt{1-z} \sqrt{\frac{1}{2}-z} \sqrt{-z} dz,$$

where  $C$  is the contour bounded by the line  $x = -\frac{N-3}{2}$  and that part of the circle  $|z| = m + \frac{1}{2}$ , where  $m$  may be increased indefinitely, which lies to the right of this line; the contour is traversed in a counter-clockwise direction.

The value of  $J = \int_C e^{\xi z} B^z \sqrt{1-z} \sqrt{\frac{1}{2}-z} \sqrt{-z} dz$  is  $2\pi i$

times the sum of the residues at the poles within the contour  $C$ .

For  $z=0$  there is a simple pole at which the residue is  $\sqrt{\frac{1}{2}} = \pi^{1/2}$

For  $z = \frac{1}{2} + \alpha$ ,  $\alpha = 0, 1, 2, \dots$ , there is a simple pole at which the

\* MacRobert,<sup>20</sup> p. 139, Rule II.

\*\* MacRobert,<sup>20</sup> p. 114 Lemma.

residue is

$$\frac{(-1)^{n+1} \pi^2 e^{\xi(n+\frac{1}{2})} B^{n+\frac{1}{2}}}{n! \sqrt{n+\frac{1}{2}} \sqrt{n+\frac{3}{2}}}$$

since

$$\sqrt{-z} = -\frac{\pi}{\sin \pi z \sqrt{1+z}}; \quad \sqrt{1-z} = \frac{\pi}{\sin \pi z \sqrt{z}}; \quad \sqrt{\frac{1}{2}-z} = \frac{\pi}{\cos \pi z \sqrt{\frac{1}{2}+z}};$$

and the residue of  $\frac{\pi}{\cos \pi z \sqrt{\frac{1}{2}+z}}$  for  $z = \frac{1}{2} + n$  is equal to  $\frac{(-1)^n}{\sqrt{n+1}}$ .

For  $z = n$  where  $n$  is an integer other than zero, the integrand has a pole of the second order, viz., that of  $\sqrt{1-z} \sqrt{-z}$  so that the residue is

$$-\pi \left[ \frac{d}{dz} \frac{e^{\xi z} B^z}{\cos \pi z \sqrt{z} \sqrt{z+\frac{1}{2}} \sqrt{z+1}} \right]_{z=n}$$

Finally we have

$$(85) \quad P(\xi) = \frac{(B e^{\xi})^{\frac{n-3}{2}}}{\sqrt{\frac{n-1}{2}} \sqrt{\frac{n-2}{2}} \sqrt{\frac{n-3}{2}}} \left\{ -\pi + \pi (e^{\xi} B)^{\frac{1}{2}} \sum_{n=0}^{\infty} \frac{(-1)^n (e^{\xi} B)^n}{n! \sqrt{n+\frac{1}{2}} \sqrt{n+\frac{3}{2}}} + \pi \sum_{n=1}^{\infty} \left[ \frac{d}{dz} \frac{(e^{\xi} B)^z}{\cos \pi z \sqrt{z} \sqrt{z+\frac{1}{2}} \sqrt{z+1}} \right]_{z=n} \right\}.$$

If we make the substitutions  $\xi = \log b = \log a N^3$  where  $a = |a_{ij}|$  and  $B = \frac{A}{N^3}$  where  $A = |A_{ijk}|$  we have for the distribution of  $a$

$$(86) \mathcal{D}(a) da = \frac{da}{a} \frac{(aA)^{\frac{N-3}{2}}}{\sqrt{\frac{N-1}{2}} \sqrt{\frac{N-2}{2}} \sqrt{\frac{N-3}{2}}} \left\{ -\frac{1}{\pi} + \frac{1}{\pi} (aA)^{\frac{1}{2}} \sum_{\lambda=0}^{\infty} \frac{(-1)^{\lambda} (aA)^{\lambda}}{\lambda! \sqrt{\frac{N-1}{2}} \sqrt{\frac{N-2}{2}} \sqrt{\frac{N-3}{2}}} + \pi \sum_{\lambda=1}^{\infty} \left[ \frac{d}{dz} \frac{(aA)^{\frac{z}{2}}}{\cos \pi z \sqrt{z} \sqrt{z+\frac{1}{2}} \sqrt{z+1}} \right]_{z=\lambda} \right\}$$

or

$$(87) \mathcal{D}(a) = \frac{A^{\frac{N-3}{2}} a^{\frac{N-5}{2}}}{\sqrt{\frac{N-1}{2}} \sqrt{\frac{N-2}{2}} \sqrt{\frac{N-3}{2}}} \left\{ -\frac{1}{\pi} + \frac{1}{\pi} (aA)^{\frac{1}{2}} \sum_{\lambda=0}^{\infty} \frac{(-1)^{\lambda} (aA)^{\lambda}}{\lambda! \sqrt{\frac{N-1}{2}} \sqrt{\frac{N-2}{2}} \sqrt{\frac{N-3}{2}}} + \pi \sum_{\lambda=1}^{\infty} \left[ \frac{d}{dz} \frac{(aA)^{\frac{z}{2}}}{\cos \pi z \sqrt{z} \sqrt{z+\frac{1}{2}} \sqrt{z+1}} \right]_{z=\lambda} \right\}$$

Case 2,  $n=4$  : With the same notation as before, we find that

$$(88) P(\xi) = \frac{1}{\sqrt{\frac{N-1}{2}} \sqrt{\frac{N-2}{2}} \sqrt{\frac{N-3}{2}} \sqrt{\frac{N-4}{2}} 2\pi} \int_{-\infty}^{\infty} e^{\frac{-it\xi}{\beta}} \frac{-it}{\beta} \sqrt{\frac{N-1}{2}+it} \sqrt{\frac{N-2}{2}+it} \sqrt{\frac{N-3}{2}+it} \sqrt{\frac{N-4}{2}+it} dt.$$

Let  $\frac{N-4}{2} + it = -z$  so that

$$(89) P(\xi) = \frac{(Be^{\frac{\xi}{\beta}})^{\frac{N-4}{2}}}{\sqrt{\frac{N-1}{2}} \sqrt{\frac{N-2}{2}} \sqrt{\frac{N-3}{2}} \sqrt{\frac{N-4}{2}} 2\pi} \int_{-\frac{N-4}{2}-i\infty}^{-\frac{N-4}{2}+i\infty} e^{\frac{\xi}{\beta} z} \frac{z}{\beta} \sqrt{\frac{3}{2}-z} \sqrt{1-z} \sqrt{\frac{1}{2}-z} \sqrt{-z} dz.$$

A similar discussion as for the case  $n=3$  applies here with regard to the convergence and we can write here too,

$$(90) P(\xi) = - \frac{(Be^{\frac{\xi}{\beta}})^{\frac{N-4}{2}}}{\sqrt{\frac{N-1}{2}} \sqrt{\frac{N-2}{2}} \sqrt{\frac{N-3}{2}} \sqrt{\frac{N-4}{2}} \cdot 2\pi i} \int_C e^{\frac{\xi}{\beta} z} \frac{z}{\beta} \sqrt{\frac{3}{2}-z} \sqrt{1-z} \sqrt{\frac{1}{2}-z} \sqrt{-z} dz,$$

where the contour  $C$  is bounded by the line  $x = -\frac{N-4}{2}$ , ( $N > 4$ ) and that part of the circle  $|z| = m + \frac{1}{2}$ , where  $m$  may be increased indefinitely, which lies to the right of this line. The contour is traversed in a counter-clockwise direction.

The value of  $J = \int_C e^{\frac{\xi}{2} z} B^z \sqrt{\frac{3}{2}-z} \sqrt{1-z} \sqrt{\frac{1}{2}-z} \sqrt{-z} dz$

is  $2\pi i$  times the sum of the residues at the poles within this contour.

For  $z=0$  there is a simple pole at which the residue is  $\sqrt{\frac{3}{2}} \sqrt{\frac{1}{2}} = \frac{\pi}{2}$ .

For  $z = \frac{1}{2}$  there is a simple pole at which the residue is  $-2\pi (e^{\frac{\xi}{2}} B)^{1/2}$ .

The integrand may also be written as

$$\frac{\pi^2}{\sin^2 \pi z} \cdot \frac{\pi i}{\cos^2 \pi z} \cdot \frac{e^{\frac{\xi}{2} z} B^z}{\sqrt{z+1} \sqrt{z+\frac{1}{2}} \sqrt{z} \sqrt{z-\frac{1}{2}}}$$

and the poles are those of  $\frac{1}{\sin^2 \pi z}$  and  $\frac{1}{\cos^2 \pi z}$ .

We have already considered the simple poles  $z=0, \frac{1}{2}$

For  $z = n$ ,  $n$  an integer other than zero, the integrand has a pole of the second order, that of  $\frac{1}{\sin^2 \pi z}$  at which the residue is

$$\pi^2 \left[ \frac{d}{dz} \frac{(e^{\frac{\xi}{2}} B)^z}{\cos^2 \pi z \sqrt{z+1} \sqrt{z+\frac{1}{2}} \sqrt{z} \sqrt{z-\frac{1}{2}}} \right]_{z=n}$$

For  $z = \frac{1}{2} + n$ ,  $n$  an integer other than zero, the integrand has a pole of the second order, that of  $\frac{1}{\cos^2 \pi z}$  at which the residue is

$$\pi^2 \left[ \frac{d}{dz} \frac{(e^{\frac{\xi}{2}} B)^z}{\sin^2 \pi z \sqrt{z+1} \sqrt{z+\frac{1}{2}} \sqrt{z} \sqrt{z-\frac{1}{2}}} \right]_{z=\frac{1}{2}+n}$$



We thus find that

$$(91) \quad P(\xi) = - \frac{(e^{\frac{\xi}{2}} B)^{\frac{N-4}{2}}}{\sqrt{\frac{N-1}{2}} \sqrt{\frac{N-2}{2}} \sqrt{\frac{N-3}{2}} \sqrt{\frac{N-4}{2}}} \left\{ \frac{\pi}{2} - 2\pi (e^{\frac{\xi}{2}} B)^{1/2} \right. \\ \left. + \pi^2 \int_{\lambda=1}^{\infty} \left[ \frac{d}{dz} \frac{(e^{\frac{\xi}{2}} B)^z}{\cos^2 \pi z \sqrt{z+1} \sqrt{z+\frac{1}{2}} \sqrt{z} \sqrt{z-\frac{1}{2}}} \right]_{z=\lambda} + \pi^2 \int_{\lambda=1}^{\infty} \left[ \frac{d}{dz} \frac{(e^{\frac{\xi}{2}} B)^z}{\sin^2 \pi z \sqrt{z+1} \sqrt{z+\frac{1}{2}} \sqrt{z} \sqrt{z-\frac{1}{2}}} \right]_{z=\frac{1}{2}+\lambda} \right\}.$$

For the distribution of  $a$ , we find

$$(92) \quad D(a) = \frac{A^{\frac{N-4}{2}} a^{\frac{N-6}{2}}}{\sqrt{\frac{N-1}{2}} \sqrt{\frac{N-2}{2}} \sqrt{\frac{N-3}{2}} \sqrt{\frac{N-4}{2}}} \left\{ -\frac{\pi}{2} + 2\pi (aA)^{1/2} \right. \\ \left. - \pi^2 \int_{\lambda=1}^{\infty} \left[ \frac{d}{dz} \frac{(aA)^z}{\cos^2 \pi z \sqrt{z+1} \sqrt{z+\frac{1}{2}} \sqrt{z} \sqrt{z-\frac{1}{2}}} \right]_{z=\lambda} - \pi^2 \int_{\lambda=1}^{\infty} \left[ \frac{d}{dz} \frac{(aA)^z}{\sin^2 \pi z \sqrt{z+1} \sqrt{z+\frac{1}{2}} \sqrt{z} \sqrt{z-\frac{1}{2}}} \right]_{z=\frac{1}{2}+\lambda} \right\}.$$

Case 3,  $n$  even: As is evident from the previous discussion, retaining the same notation,

$$(93) \quad P(\xi) = \frac{1}{\prod_{j=1}^n \sqrt{\frac{N-j}{2}} 2\pi} \int_{-\infty}^{\infty} e^{-it\xi} B^{\frac{n}{2}} \prod_{j=1}^n \sqrt{\frac{N-j}{2} + it} dt.$$

Let  $\frac{N-n}{2} + i t = -z$  so that

$$(94) \quad P(\xi) = \frac{(e^{\xi} B)^{\frac{N-n}{2}}}{\prod_{j=1}^n \sqrt{\frac{N-j}{2}} 2\pi i} \int_{-\frac{N-n}{2}-i\infty}^{-\frac{N-n}{2}+i\infty} e^{\xi z} B^z \sqrt{\frac{n-1}{2}-z} \sqrt{\frac{n-2}{2}-z} \cdots \sqrt{-z} dz.$$

The same considerations as to the convergence and the contour are applicable here too and we find that

$$(95) \quad P(\xi) = - \frac{(e^{\xi} B)^{\frac{N-n}{2}}}{\prod_{j=1}^n \sqrt{\frac{N-j}{2}} 2\pi i} \int_C e^{\xi z} B^z \sqrt{\frac{n-1}{2}-z} \sqrt{\frac{n-2}{2}-z} \cdots \sqrt{-z} dz,$$

where  $C$  is the contour bounded by the line  $\chi = -\frac{N-n}{2}$ , ( $N > n$ ) and that part of the circle  $|z| = n + \frac{1}{2}$ , where  $n$  may increase indefinitely, to the right of this line and the contour is traversed in a counter-clockwise direction.

$$\text{The value of } \int_C e^{\xi z} B^z \sqrt{\frac{n-1}{2}-z} \sqrt{\frac{n-2}{2}-z} \cdots \sqrt{-z} dz$$

is  $2\pi i$  times the sum of the residues at the poles within the contour. Let us write  $n = 2p$  so that the integrand is

$$e^{\xi z} B^z \sqrt{\frac{2p-1}{2}-z} \sqrt{\frac{2p-2}{2}-z} \cdots \sqrt{-z}.$$

For  $z = \lambda$ ,  $\lambda = 0, 1, 2, \dots, p-2$  there is a pole of the  $(\lambda+1)^{th}$  order, the integrand being representable in the form

$$(96) \quad \frac{\pi^{1+2+\cdots+\lambda+1} e^{\xi z} B^z \sqrt{\frac{1}{2}-z} \sqrt{\frac{3}{2}-z} \cdots \sqrt{\frac{2p-1}{2}-z} \sqrt{\lambda+1-z} \sqrt{\lambda+2-z} \cdots \sqrt{p-1-z}}{\sin \pi \lambda \pi z \sqrt{z+1} \sqrt{z} \cdots \sqrt{z-\lambda+1}}.$$

The residue is therefore,

$$\frac{(-1)^{\frac{n(n+1)}{2}}}{n(n+1)} \left[ \frac{d^n}{dz^n} e^{\xi z} B^z \frac{z}{\sqrt{\frac{1}{2}-z} \sqrt{\frac{3}{2}-z} \cdots \sqrt{\frac{2p-1}{2}-z} \sqrt{n+1-z} \sqrt{n+2-z} \cdots \sqrt{p-1-z}} \right]_{z=n}$$

For  $z = \frac{1}{2} + n$ ,  $n = 0, 1, 2, \dots, p-2$  there is a pole of the order, the integrand being representable in the form

$$\frac{(-1)^{0+1+2+\dots+n}}{\pi} e^{\xi z} B^z \frac{z}{\cos^{\frac{n+1}{2}} \pi z \sqrt{z+\frac{1}{2}} \sqrt{z-\frac{1}{2}} \cdots \sqrt{z-n+\frac{1}{2}}}$$

The residue is therefore of the form

$$\frac{(-1)^{\frac{n(n+1)}{2}}}{(n+1)(n+1)} \left[ \frac{d^n}{dz^n} e^{\xi z} B^z \frac{z}{\sqrt{z+\frac{1}{2}} \sqrt{z-\frac{1}{2}} \cdots \sqrt{z-n+\frac{1}{2}}} \right]_{z=\frac{1}{2}+n}$$

For  $z = p-1+n$ ,  $n = 0, 1, 2, \dots$  there is a pole of the order, the integrand being representable as

$$\frac{(-1)^p \pi^p e^{\xi z} B^z}{\sin^p \pi z \cos^p \pi z \sqrt{z+1} \sqrt{z+\frac{1}{2}} \cdots \sqrt{z-p+\frac{3}{2}}}$$

The residue is therefore of the form

$$\frac{(-1)^p \pi^p}{(p-1)!} \left[ \frac{d^{p-1}}{dz^{p-1}} \frac{e^{\xi z} B^z}{\cos^p \pi z \sqrt{z+1} \sqrt{z+\frac{1}{2}} \cdots \sqrt{z-p+\frac{3}{2}}} \right]_{z=p-1+n}$$

For  $z = \frac{2p-1}{2} + n$ ,  $n = 0, 1, 2, \dots$  there is a pole of the order at which the residue is

$$\frac{(-1)^p \pi^p}{(p-1)!} \left[ \frac{d^{p-1}}{dz^{p-1}} \frac{e^{\xi z} B^z}{\sin^p \pi z \sqrt{z+1} \sqrt{z+\frac{1}{2}} \cdots \sqrt{z-p+\frac{3}{2}}} \right]_{z=\frac{2p-1}{2}+n}$$

We have therefore that

$$(96) \quad P(\xi) =$$

$$\begin{aligned} & \frac{(Be^{\xi})^{\frac{N-n}{2}}}{\prod_{j=1}^n \sqrt{\frac{N-j}{2}}} \left\{ \sum_{\lambda=0}^{p-2} \frac{(-1)^{\frac{(\lambda+1)(\lambda+2)}{2}}}{\lambda!} \left[ \frac{d^{\lambda}}{dz^{\lambda}} \frac{e^{\xi z} B^{\frac{z}{2}} \sqrt{\frac{1}{2}-z} \sqrt{\frac{3}{2}-z} \cdots \sqrt{\frac{2p-1}{2}-z} \sqrt{\lambda+1-z} \sqrt{\lambda+2-z} \cdots \sqrt{p-1-z}}{\sqrt{z+1} \sqrt{z} \cdots \sqrt{z-\lambda+1}} \right]_{z=\lambda} \right. \\ & + \sum_{\lambda=0}^{p-2} \frac{(-1)^{\frac{(\lambda+1)(\lambda+2)}{2}}}{\lambda!} \left[ \frac{d^{\lambda}}{dz^{\lambda}} \frac{e^{\xi z} B^{\frac{z}{2}} \sqrt{-z} \sqrt{1-z} \cdots \sqrt{p-1-z} \sqrt{\lambda+\frac{1}{2}-z} \sqrt{\lambda+\frac{3}{2}-z} \cdots \sqrt{\frac{2p-1}{2}-z}}{\sqrt{z+\frac{1}{2}} \sqrt{z-\frac{1}{2}} \cdots \sqrt{z-\lambda+\frac{1}{2}}} \right]_{z=\frac{1}{2}+\lambda} \\ & + \sum_{\lambda=0}^{\infty} \frac{(-1)^{\frac{p(\lambda+p)}{2}}}{(p-1)!} \left[ \frac{d^{p-1}}{dz^{p-1}} \frac{e^{\xi z} B^{\frac{z}{2}}}{\cos^p \pi z \sqrt{z+1} \sqrt{z+\frac{1}{2}} \cdots \sqrt{z-p+\frac{1}{2}}} \right]_{z=p-1+\lambda} \\ & + \sum_{\lambda=0}^{\infty} \frac{(-1)^{\frac{p(\lambda+p+1)}{2}}}{(p-1)!} \left[ \frac{d^{p-1}}{dz^{p-1}} \frac{e^{\xi z} B^{\frac{z}{2}}}{\sin^p \pi z \sqrt{z+1} \sqrt{z+\frac{1}{2}} \cdots \sqrt{z-p+\frac{1}{2}}} \right]_{z=\frac{2p-1}{2}+\lambda} \end{aligned}$$

For the distribution of  $a$ , we find

$$(97) \quad D(a) =$$

$$\frac{A a^{\frac{N-n}{2}}}{\prod_{j=1}^n \sqrt{\frac{N-j}{2}}} \left\{ \sum_{\lambda=0}^{\frac{n(n-1)}{2}} \frac{(-1)^{\frac{\lambda(\lambda-1)}{2}}}{\lambda!} \left[ \frac{d^{\lambda}}{dz^{\lambda}} \frac{(aA)^{\frac{z}{2}} \sqrt{\frac{1}{2}-z} \sqrt{\frac{3}{2}-z} \cdots \sqrt{\frac{2p-1}{2}-z} \sqrt{\lambda+1-z} \sqrt{\lambda+2-z} \cdots \sqrt{p-1-z}}{\sqrt{z+1} \sqrt{z} \cdots \sqrt{z-\lambda+1}} \right]_{z=\lambda} \right.$$

$$\begin{aligned}
& + \left[ \frac{(-1)^{\frac{n(n+3)}{2}}}{n!} \left[ \frac{d^n}{dz^n} \frac{(aA)^z}{\sqrt{z+\frac{1}{2}} \sqrt{z-\frac{1}{2}} \cdots \sqrt{z-n+\frac{1}{2}}} \right]_{z=\frac{1}{2}+n} \right. \\
& + \left[ \frac{(-1)^{p(n+p)+1}}{(p-1)!} \left[ \frac{d^{p-1}}{dz^{p-1}} \frac{(aA)^z}{\cos^p \pi z \sqrt{z+1} \sqrt{z+\frac{1}{2}} \cdots \sqrt{z-p+\frac{3}{2}}} \right]_{z=p-1+n} \right. \\
& + \left. \left[ \frac{(-1)^{p(p+n+1)+1}}{(p-1)!} \left[ \frac{d^{p-1}}{dz^{p-1}} \frac{(aA)^z}{\sin^p \pi z \sqrt{z+1} \sqrt{z+\frac{1}{2}} \cdots \sqrt{z-p+\frac{3}{2}}} \right]_{z=\frac{2p-1}{2}+n} \right]
\end{aligned}$$

with  $n=2p$ .

Case 4,  $n$  odd: As before we find that

$$(98) \quad P(\xi) = - \frac{(e^{\xi} B)^{\frac{N-n}{2}}}{\prod_{j=1}^p \sqrt{\frac{N-j}{2}} 2\pi i} \int_C e^{\xi z} B^z \sqrt{\frac{n-1}{2}-z} \sqrt{\frac{n-3}{2}-z} \cdots \sqrt{-z} dz.$$

Let  $n=2p+1$

The integrand is

$$e^{\xi z} B^z \sqrt{p-z} \sqrt{\frac{2p-1}{2}-z} \cdots \sqrt{\frac{1}{2}-z} \sqrt{-z}.$$

The considerations are similar to the case for  $n$  even except that the integrand has an additional factor, viz.  $\sqrt{p-z}$ .

For  $z=n$ ,  $n=0, 1, 2, \dots, (p-1)$  there is a pole of the  $(n+1)-th$  order at which the residue is

$$\frac{(-1)^{\frac{(n+1)(n+2)}{2}}}{(-1)^{\frac{n(n+1)}{2}} n!} \left[ \frac{d^n}{dz^n} \frac{e^{\xi z} B^z \sqrt{\frac{1}{2}-z} \sqrt{\frac{3}{2}-z} \cdots \sqrt{\frac{2p-1}{2}-z} \sqrt{n+1-z} \sqrt{n+2-z} \cdots \sqrt{p-z}}{\sqrt{z+1} \sqrt{z} \cdots \sqrt{z-n+1}} \right]_{z=n}$$

For  $z = \frac{1}{2} + n$ ,  $n = 0, 1, \dots, (p-2)$  there is a pole of the  $(n+1)-th$  order at which the residue is

$$\frac{(-1)^{\frac{n(n+1)}{2}}}{(n+1)(n+1)!} \left[ \frac{d^n}{dz^n} \frac{e^{\xi z} B^z}{\sqrt{z+\frac{1}{2}} \sqrt{z-\frac{1}{2}} \dots \sqrt{p-\frac{1}{2}} \sqrt{n+\frac{1}{2}-z} \dots \sqrt{\frac{z-p-1}{2}-z}} \right]_{z=\frac{1}{2}+n}$$

For  $z = p+n$ ,  $n = 0, 1, 2, \dots$  there is a pole of the  $(p+1)-th$  order at which the residue is

$$\frac{(-1)^{p+1} \pi^p}{(p+1)(p+1)!} \left[ \frac{d^p}{dz^p} \frac{e^{\xi z} B^z}{\cos^p \pi z \sqrt{z+1} \sqrt{z+\frac{1}{2}} \dots \sqrt{z-p+1}} \right]_{z=p+n}$$

For  $z = \frac{2p-1}{2} + n$ ,  $n = 0, 1, 2, \dots$  there is a pole of the  $p-th$  order at which the residue is

$$\frac{(-1)^{p+1} \pi^{p+1}}{(p+1)(p+1)!} \left[ \frac{d^{p-1}}{dz^{p-1}} \frac{e^{\xi z} B^z}{\sin^{p+1} \pi z \sqrt{z+1} \sqrt{z+\frac{1}{2}} \dots \sqrt{z-p+1}} \right]_{z=\frac{2p-1}{2}+n}$$

since the integrand is representable as

$$\frac{(-1)^{p+1} \pi^{p+1} \pi^p e^{\xi z} B^z}{\sin^{p+1} \pi z \cdot \cos^p \pi z \sqrt{z+1} \sqrt{z+\frac{1}{2}} \dots \sqrt{z-p+1}}$$

We have therefore that

$$(99) \quad P(\xi) =$$

$$\begin{aligned} & \frac{(e^{\xi})^{\frac{N-n}{2}}}{\prod_{j=1}^n \sqrt{\frac{N-j}{2}}} \left\{ \sum_{n=0}^{p-1} \frac{(-1)^{\frac{n(n+1)(2n+2)}{2}}}{n!} \left[ \frac{d^n}{dz^n} \frac{e^{\xi z} B^z}{\sqrt{z+\frac{1}{2}} \sqrt{z-\frac{1}{2}} \dots \sqrt{p-\frac{1}{2}} \sqrt{n+\frac{1}{2}-z} \dots \sqrt{\frac{z-p-1}{2}-z}} \right]_{z=\frac{1}{2}+n} \right. \\ & \left. + \sum_{n=0}^{p-2} \frac{(-1)^{\frac{n(n+1)(2n+2)}{2}}}{n!} \left[ \frac{d^n}{dz^n} \frac{e^{\xi z} B^z}{\sqrt{z+\frac{1}{2}} \sqrt{z-\frac{1}{2}} \dots \sqrt{z-n+\frac{1}{2}}} \right]_{z=n+\frac{1}{2}} \right\} \end{aligned}$$

$$\begin{aligned}
& + \left\{ \sum_{n=0}^{\infty} \frac{(-1)^n \pi^{p+n}}{p!} \left[ \frac{d^p}{dz^p} \frac{e^{\xi z} B^z}{\cos^p \pi z \sqrt{z+1} \sqrt{z+\frac{1}{2}} \cdots \sqrt{z-p+1}} \right]_{z=p+n} \right. \\
& \left. + \sum_{n=0}^{\infty} \frac{(-1)^n \pi^{p+n+1}}{(p-1)!} \left[ \frac{d^{p-1}}{dz^{p-1}} \frac{e^{\xi z} B^z}{\sin^{p+1} \pi z \sqrt{z+1} \sqrt{z+\frac{1}{2}} \cdots \sqrt{z-p+1}} \right]_{z=\frac{2p-1}{2}+n} \right\}
\end{aligned}$$

The distribution for  $a$  is

$$(100) \quad D(a) =$$

$$\begin{aligned}
& \frac{A^{\frac{N-p}{2}} a^{\frac{N-p-2}{2}}}{\prod_{j=1}^p \sqrt{\frac{N-j}{2}}} \left\{ \sum_{n=0}^{p-1} \frac{(-1)^n \pi^{p-n}}{n!} \left[ \frac{d^n}{dz^n} \frac{(aA)^z \sqrt{\frac{1}{2}-z} \sqrt{\frac{3}{2}-z} \cdots \sqrt{\frac{2p-1}{2}-z} \sqrt{z+1} \sqrt{z+\frac{1}{2}} \cdots \sqrt{z-p+1}} \right]_{z=n} \right. \\
& + \sum_{n=0}^{p-2} \frac{(-1)^n \pi^{p-n-1}}{n!} \left[ \frac{d^n}{dz^n} \frac{(aA)^z \sqrt{-z} \sqrt{1-z} \cdots \sqrt{p-z} \sqrt{z+\frac{1}{2}} \sqrt{z+\frac{3}{2}} \cdots \sqrt{\frac{2p-1}{2}-z}} \right]_{z=n+\frac{1}{2}} \\
& + \sum_{n=0}^{\infty} \frac{(-1)^n \pi^{p+n}}{p!} \left[ \frac{d^p}{dz^p} \frac{(aA)^z}{\cos^p \pi z \sqrt{z+1} \sqrt{z+\frac{1}{2}} \cdots \sqrt{z-p+1}} \right]_{z=p+n} \\
& \left. + \sum_{n=0}^{\infty} \frac{(-1)^n \pi^{p+n+1}}{(p-1)!} \left[ \frac{d^{p-1}}{dz^{p-1}} \frac{(aA)^z}{\sin^{p+1} \pi z \sqrt{z+1} \sqrt{z+\frac{1}{2}} \cdots \sqrt{z-p+1}} \right]_{z=\frac{2p-1}{2}+n} \right\},
\end{aligned}$$

with  $n = 2p+1$ .

It is of interest to derive from the general formula the distribution when  $n = 1, 2$ .

For  $n = 1$  the value of  $p$  in equation (100) is zero. The expression in the brace in equation (100) becomes

$$1 - \frac{aA}{1!} + \frac{(aA)^2}{2!} - \dots = e^{-aA},$$

so that

$$(101) \quad D(a) = \frac{A^{\frac{N-1}{2}} a^{\frac{N-3}{2}} e^{-aA}}{\sqrt{\frac{N-1}{2}}}.$$

For  $N = 2$  the value of  $p$  in equation (97) is 1. The expression in the brace in equation (97) becomes

$$\begin{aligned} (102) \quad & \frac{\pi}{\sqrt{\frac{1}{2}}} + \frac{\pi aA}{\sqrt{2} \sqrt{\frac{3}{2}}} + \frac{\pi (aA)^2}{\sqrt{3} \sqrt{\frac{5}{2}}} + \dots \\ & - \frac{\pi (aA)^{1/2}}{\sqrt{\frac{3}{2}} \sqrt{1}} - \frac{\pi (aA)^{3/2}}{\sqrt{\frac{5}{2}} \sqrt{2}} - \frac{\pi (aA)^{5/2}}{\sqrt{\frac{7}{2}} \sqrt{3}} - \dots \\ & = \frac{\pi}{\sqrt{\frac{1}{2}}} \left[ 1 - \frac{2(aA)^{1/2}}{1!} + \frac{2^2 aA}{2!} - \frac{2^3 (aA)^{3/2}}{3!} + \dots \right] \\ & = \pi^{1/2} e^{-2\sqrt{aA}}; \end{aligned}$$

there is no difficulty about combining the infinite series in equation (102) since each is absolutely convergent for all value of  $a$ .



Therefore,

$$(103) \quad D(a) = \frac{\pi^{1/2} A^{\frac{N-2}{2}} a^{\frac{N-4}{2}} e^{-2\sqrt{a}A}}{\sqrt{\frac{N-1}{2}} \sqrt{\frac{N-2}{2}}} = \frac{2^{N-3} A^{\frac{N-2}{2}} a^{\frac{N-4}{2}} e^{-2\sqrt{a}A}}{\sqrt{N-2}}.$$

The explicit expressions for  $N = 1, 2$  have already been obtained otherwise by Wilks.<sup>41</sup>

### PART 3

#### Conclusion

*XIV. Summary and Conclusions.* By the use of a discontinuity factor derived from Fourier's Integral Theorem we obtain the characteristic function (in the sense of P. Levy) of the distribution law, and the distribution law of very general functions of variables satisfying a continuous distribution law. In the application of the general theory a certain lemma is found to simplify the calculations for a particular class of distribution laws and functions. Several of the distributions derived are presented not because the results are new but as illustrations of a general method of procedure which it is hoped will enable us to find the distribution laws of many functions not yet obtained.

The explicit form of the distribution of the generalized sample variance for an  $n$ -variate normal population is derived. The same analysis is applicable to find the explicit form of the other generalizations introduced by Wilks, for general  $n$ , since the integrals that must be evaluated are all of the same general nature. The writer hopes to be able to present these further results in the near future.

#### NOTE

After this paper had been completed, the writer's attention was drawn to the fact that an analysis very similar to that of Sections VIII, X, and XI

of this paper had already appeared in two papers by Wishart and Bartlett, viz:

"The distribution of second order moment statistics in a normal system." Proc. Cambridge Phil. Soc. Vol. 28 (1932) p. 455f.

"The generalized product moment distribution in a normal system." Proc. Cambridge Phil. Soc. Vol. 29 (1933) p. 260.

These sections are, however, presented here as illustrations of the Lemma of section VII.

#### BIBLIOGRAPHY

1. Bôcher: Introduction to Higher Algebra, pp. 30-33.
2. Cauchy: Comptes Rendus, Vol. 37 (1853), pp. 100, 150, 198, 264, 326.
3. Charlier, C. V. L.: Arkiv Fur Math. Astron. Och Fysik. Vol. 2 (1905-6) No. 8, No. 15; Vol. 4 (1908) No. 13; Vol. 5 (1909) No. 15; Vol. 7 (1912) No. 17; Vol. 8 (1912) No. 2, No. 4; Vol. 9 (1913) No. 25, No. 26.
4. Czuber, E.: Wahrscheinlichkeitsrechnung, I (1914), p. 66.
5. Dodd, E. L.: The Frequency Law of a Function of One Variable. Bull. Am. Math. Soc., Vol. 31 (1925) p. 27.
6. Dodd, E. L.: The Frequency Law of a Function of Variables with Given Frequency Laws, Annals of Math., 2nd S., Vol. 27 (1925), pp. 12-20.
7. Craig, A. T.: On the Distribution of Certain Statistics, Am. Jour. of Math., Vol. LIV (1932), pp. 353-366.
8. Fisher, R. A.: Frequency Distribution of the Values of the Correlation Coefficient in Samples from an indefinitely Large Population, Biometrika, Vol. 10 (1914-15), pp. 507-21.
9. Fisher, R. A.: On the Interpretation of  $\chi^2$  from Contingency Tables and the Calculation of  $P$ ; Jour. Roy. Stat. Soc., Vol. 85, p. 87.
10. Gronwall, T. H.: The Theory of the Gamma Function; Annals of Math., Vol. 20 (1918-19), p. 48, Th. XIII.
11. Hausdorff, F.: Beiträge zur Wahrscheinlichkeitsrechnung. Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig. Berichte über die Verhandlungen Math.—Phys. Classe, Vol. 53 (1901, pp. 152-178.
12. Hobson: The Theory of Functions of a Real Variable (1907), p. 590.
13. Irwin, J. O.: On the Frequency Distribution of the Means of Samples from a Population having any Law of Frequency with Finite Moments with Special Reference to Pearson's Type II; Biometrika, Vol. 19 (1927), pp. 225-39.
14. Kameda, T.: Theorie der erzeugenden funktion und ihre anwendung auf die Wahrscheinlichkeits-Rechnung, Proc. Math. Phys. Soc.; Tokyo, Vol. 8 (1915-16), pp. 262, 336, 556 ff.
15. Kameda, T.: Eine Verallgemeinerung des Poissonschen Problems in der Wahrscheinlichkeits-Rechnung; Proc. Math. Phys. Soc., Tokyo, Vol. 9 (1917-18), pp. 155 ff.
16. Laplace: Théorie Analytique des Probabilités, 3rd Ed. (1820), pp. 3 ff; pp. 80 ff.

17. Lévy, P.: Calcul des Probabilités, p. 161.
18. Lévy, P.: Comptes Rendus, Vol. 176 (1923), pp. 1118-1120; pp. 1284-1286.
19. Lévy, P.: Bull. de la Soc. Math. de France, Vol. 52 (1924), pp. 49-85.
20. MacRobert, T. M.: Functions of a Complex Variable (1925).
21. Molina, E. C.: The Theory of Probability: Some Comments on Laplace's Théorie Analytique; Bull. Am. Math. Soc.; Vol. 36 (1930), pp. 369 ff.
22. Pearson, K.: On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it can be reasonably supposed to have arisen from random sampling.—Phil. Mag., 5th series, Vol. 50 (1900), p. 157.
23. Pearson, K.: On the Distribution of the Standard Deviations of Small Samples: Appendix I to papers by "Student" and R. A. Fisher, Biometrika, Vol. 10 (1914-15), pp. 522-29.
24. Pearson, K.: On a Brief Proof of the Fundamental Formula for Testing the Goodness of Fit of Frequency Distributions and on the Probable Error of P. Phil. Mag., 6th Series, Vol. 31 (1916), p. 369.
25. Pearson, K.; Jeffery, G. B.; Elderton, E. M. and F. R. S. On the Distribution of the First Product Moment Coefficient in Samples Drawn from an Indefinitely Large Normal Population, Biometrika, Vol. 21 (1929), pp. 164-201.
26. Pearson, K.; Stouffer, S. A., and David, F. N. Further Applications in Statistics of the  $J_m(x)$  Bessel Function. Biometrika, Vol. 24 (1932), pp. 293-350.
27. Poincaré, H.: Calcul des Probabilités, 2nd Ed. (1923), p. 206.
28. Poisson: Connaissance des temps de l'année, 1827.
29. Poisson: Recherches sur la Prob. Chap. IV.
30. Rhodes, E. C.: On the Problem Whether two given Samples can be Supposed to have been drawn from the Same Population. Biometrika, Vol. 16 (1924), p. 239.
31. Rider, P. R.: A Survey of the Theory of Small Samples; Annals of Math., 2nd S., Vol. 31 (1930), pp. 577-628.
32. Rietz, H. L.: On a Certain Law of Probability of Laplace; International Math. Congress, Toronto, Canada, 1924.
33. Rietz, H. L.: On the Representation of a Certain Fundamental Law of Probability; Trans. Am. Math. Soc., Vol. 27 (1925), pp. 197-212.
34. Romanovsky, V.: On the Moments of Standard Deviation and of Correlation Coefficient in Samples from Normal. Metron., Vol. 5 (1925), No. 4, pp. 3-46.
35. Schols, Ch. M.: Demonstration directe de la loi limite pour les erreurs dans le plan et dans l'espace. Annals d'Ecole Polytechnique de Delft. Vol. 3 (1887), p. 195 ff.
36. "Student": The Probable Error of a Mean; Biometrika, Vol. 6 (1908-9), pp. 1-25.
37. Watson, G. N.: A Treatise on the Theory of the Bessel Function.

38. Webster, A. G.: Partial Differential Equations of Math. Physics (1927), p. 158 ff.
39. Whittaker & Robinson: The Calculus of Observations (1924).
40. Whittaker & Watson: Modern Analysis, 2nd Ed. (1915).
41. Wilks, S. S.: Certain Generalizations in the Analysis of Variance, *Biometrika*, Vol. 24 (1932), pp. 471-94.
42. Wishart, J.: The Generalized Product Moment Distribution in Samples from a Normal Multivariate Population. *Biometrika*, Vol. XXA (1928), pp. 32-52.

# ON MEASURES OF CONTINGENCY

By

FRANK M. WEIDA

1. *Introduction.* When we deal with the problem of relationship of attributes, we may classify each attribute into a number of groups. To illustrate: If the attributes are  $x_i$  ( $i=1, 2, 3, \dots, n$ ) and if the group belonging to  $X_i$  is  $x_i^j$  ( $j=1, 2, 3, \dots, m_i$ ), that belonging to  $X_2$  is  $x_2^j$  ( $j=1, 2, 3, \dots, m_2$ ), ..., that belonging to  $X_i$  is  $x_i^k$  ( $k=1, 2, 3, \dots, m_i$ ), ..., we may form an  $m_1 \times m_2 \times \dots \times m_i \times \dots$  table which contains  $m_1 \times m_2 \times \dots \times m_i \times \dots$  compartments. In this fashion, it is possible to distribute the total frequency of the "universe" or the "sub-universe" into sub-groups which correspond to these  $m_1 \times m_2 \times \dots \times m_i \times \dots$  compartments.

For such situations, Pearson<sup>1</sup> and others<sup>2</sup> have suggested certain measures of relation between the attributes. We shall in this paper be interested primarily in Pearson's measures of contingency. In the case of two attributes, Pearson proceeds as follows: Suppose that  $A$  is any attribute and let it be classified into the groups  $A_i$  ( $i=1, 2, 3, \dots, s$ ) and let  $B$  be another attribute classified into the groups  $B_j$  ( $j=1, 2, 3, \dots, t$ ). Let the total number of individuals examined be  $N$ . Now, the probability a-priori of an individual falling into the respective groups  $A_i$  is  $n_i/N$  where  $n_i$  is the number which fall into  $A_i$ . Again, if  $m_j$  is the number which fall into  $B_j$ , then the probability a-priori of an individual falling into the respective groups  $B_j$  is  $m_j/N$  where  $m_j$  is the number which fall into  $B_j$ . If the attributes are independent in the probability sense, then, if  $N$  pairs of attri-

<sup>1</sup> Pearson, Karl, "On the Theory of Contingency and its Relation to Association and Normal Correlation," *Drapers' Company Research Memoirs, Biometric Series i.*; Dulau & Co., London, 1904.

<sup>2</sup> Yule, G. Udny, "An Introduction to the Theory of Statistics," Charles Griffin & Company, Limited, London, 1927, pp. 17-74.

butes are examined, the number expected in the  $(i,j)$  compartment is

$$N \cdot \frac{n_{i.}}{N} \cdot \frac{m_{.j}}{N} = \frac{n_{i.} m_{.j}}{N} = \nu_{ij}.$$

Suppose the number observed is  $n_{ij}$ . Then, if we allow for the errors of random sampling,  $(n_{ij} - \nu_{ij})$  is the departure from independent probability of the occurrence of the groups  $A, B$ . Then, any measure of the total departure from independent probability is termed by Pearson a measure of contingency. Consequently, the measure of contingency is some function of the  $(n_{ij} - \nu_{ij})$  quantities for the whole table.

Again, for a given

$$\chi^2 = \sum \left\{ \frac{(n_{ij} - \nu_{ij})^2}{\nu_{ij}} \right\}$$

Pearson has shown how to obtain the probability<sup>3</sup>  $P$  as a measure to determine how far the observed system is not compatible with a basis of independent probability. He calls  $(1-P)$  the *contingency grade* and

$$\phi^2 = \frac{\chi^2}{N}$$

the *mean square contingency*. Also,

$$\psi = \frac{\sum (n_{ij} - \nu_{ij})}{N}$$

is the *mean contingency* when  $\sum$  refers to summation for all positive terms.

In his theory of contingency, Pearson appears to use the definition of probability used in practically all treatises on the subject.

<sup>3</sup> Pearson, Karl, "On the criterion that a given system of deviations from the probable in the case of correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *Phil. Mag.*, Series V. 1. 157-175.

This definition excludes the whole field of statistical probability. It appears fairly obvious that the development of statistical concepts is approached more naturally from a limit definition for probability than from the familiar definitions suggested by games of chance. It is the purpose of this paper to improve the treatment of Pearson's theory of contingency and make it more elegant for theoretical as well as empirical discussions. To accomplish this we make use of the notion of *characteristic function*<sup>4</sup> and a definition of probability that includes all forms of probability. It is believed that we have thus idealized Pearson's conception of contingency. We discuss multiple as well as partial contingency. We also consider briefly the case of certain dependent events and the concept of mutual exclusiveness, as well as the concept of connection.

2. *Definitions and assumptions.* In our discussion we need and use the following definitions and assumptions:<sup>5</sup>

*Assumption I.* If an event which can happen in two different ways be repeated a great number of times under the same essential conditions, the ratio of the number of times that it happens in one way to the total number of trials, will approach a definite limit as the latter number increases indefinitely.

*Definition I.* The limit described in assumption I we call the *probability* that the event shall happen in the first way under these conditions.

*Assumption II.* If an event can happen in a certain number of ways, all of which are equally likely, and if a certain number of these be called favorable, then the ratio of the number of favorable ways to the total number is equal to the probability that the event will turn out favorably.

*Assumption III.* If an event depend on  $n$  independent varia-

<sup>4</sup> The characteristic function of  $A$  is that function which is equal to unity for the elements of  $A$  and zero elsewhere. Usually  $A$  is assumed to be a sub-class of some class on which the characteristic function is defined.

<sup>5</sup> Coolidge, J. L., "An Introduction to Mathematical Probability," The Clarendon Press, 1925, pp. 1-12.

bles  $X_1, X_2, \dots, X_n$  which can vary continuously in an  $n$ -dimensional continuous manifold, there exists such an analytic function  $F(X_1, \dots, X_n)$  that the probability for a result corresponding to a group of values in the infinitesimal region

$$X_1 \pm \frac{1}{2} dX_1, \quad X_2 \pm \frac{1}{2} dX_2, \dots, \quad X_n \pm \frac{1}{2} dX_n$$

differs by an infinitesimal of higher order from

$$F(X_1, X_2, \dots, X_n) dX_1 dX_2 \dots dX_n.$$

*Definition II.* If a variable  $X$  take the different values  $X_i$  ( $i=1, 2, \dots, n$ ) with the respective probabilities  $p_i$  ( $i=1, 2, \dots, n$ ) and these are all the possible values for that variable, then

$$\sum_{i=1}^n p_i X_i$$

is called the *mean value* of the variable  $X$ .

*Definition III.* Two variables are said to be independent if the probability that one lie close to a given value is independent of the value of the other.

3. *Pearson's mean square contingency.* Let the attributes be  $X$  and  $Y$ . Let  $\phi_{ij}^{\cdot}$  be the number of individuals having the group value  $X_j$  of  $X$  and  $Y_i$  of  $Y$ . The total number of individuals having the group value  $Y_i$  of  $Y$  is  $\phi_{i\cdot}^{\cdot}$ <sup>6</sup> and the total number of individuals having the group value  $X_j$  of  $X$  is  $\phi_{\cdot j}^{\cdot}$ . The total number of individuals examined then is  $\phi_{\cdot\cdot}^{\cdot}$ .

Now, suppose it is true that

$$(1) \quad F_{ij} = \phi_{i\cdot}^{\cdot} \phi_{\cdot j}^{\cdot} / \phi_{\cdot\cdot}^{\cdot},$$

where  $F_{ij} = \phi_{ij}^{\cdot} / \phi_{\cdot\cdot}^{\cdot}$ . Let  $\bar{F}_{i\cdot}, \bar{\phi}_{i\cdot}^{\cdot}, \bar{\phi}_{\cdot j}, \bar{\phi}_{\cdot j}^{\cdot}, \bar{\phi}_{\cdot\cdot}$  be, respectively, the *mean values* of  $F_{ij}, \phi_{ij}^{\cdot}, \phi_{i\cdot}^{\cdot}, \phi_{\cdot j}, \phi_{\cdot j}^{\cdot}$ .

<sup>6</sup> A repeated index means summation for all possible values of such repeated index.



Since, in the case of independence, the mean of the product is the product of the means,<sup>7</sup> we have

$$(2) \quad \overline{F_{ij}} = \overline{\phi_{ij}^i} \cdot \overline{\phi_{ij}^j}.$$

Now, if  $\phi_{ij}$  is the characteristic function of the observation,  $\phi_{ij}^i$  has the value unity if the event succeeds and zero if the event fails. Let  $p_{ij}$  be the probability that the event succeeds and  $q_{ij}$  the probability that the event fails. Then, the mean value  $\overline{\phi_{ij}^i}$  of  $\phi_{ij}^i$  is given by

$$(3) \quad \overline{\phi_{ij}^i} = p_{ij} \cdot 1 + q_{ij} \cdot 0 = p_{ij}.$$

Similarly,

$$(4) \quad \overline{\phi_{ij}^i} = p_{ij} \cdot 1 + q_{ij} \cdot 0 = p_{ij}.$$

$$(5) \quad \overline{\phi_{ij}^j} = p_{ij} \cdot 1 + q_{ij} \cdot 0 = p_{ij}.$$

$$(6) \quad \overline{\phi_{ij}^{ij}} = p_{ij} \cdot 1 + q_{ij} \cdot 0 = p_{ij}.$$

But  $p_{ij}^{ij} = 1$ , hence, in the case of independence,  $\overline{F_{ij}} = p_{ij}$ . Hence, from (2), (3), (4), (5), and (6), in the case of independence, we have

$$(7) \quad p_{ij} = p_{ij}^i \cdot p_{ij}^j.$$

In the case of dependence, we have that

$$(8) \quad p_{ij} = M(\phi_{ij}^i \phi_{ij}^j) \neq p_{ij}^i p_{ij}^j,$$

where  $M(\phi_{ij}^i \phi_{ij}^j)$  is the mean value of  $\phi_{ij}^i \phi_{ij}^j$ .<sup>8</sup>

The quantity  $(p_{ij} - p_{ij}^i p_{ij}^j)$  represents the departure between the mean value  $\phi_{ij}$  has and that which it should have in the case of independence.

Let us now consider the square of the departure relative to

<sup>7</sup> Coolidge, J. L., "An Introduction to Mathematical Probability," The Clarendon Press, 1925, p. 62.

<sup>8</sup> Tschuprow, A. A., "Grundbegriffe und grundprobleme der Korrelationstheorie," B. G. Teubner, Berlin, 1925, pp. 39-63.

$p_{ij}^i, p_{ij}^j$ , namely,

$$\psi_{ij}^2 = \frac{(p_{ij} - p_{ij}^i p_{ij}^j)^2}{p_{ij}^i p_{ij}^j}.$$

For all cases, we have

$$(9) \quad \Phi^2 = (\psi_{ij}^2)^{ij},$$

which is Pearson's *mean square contingency* and  $\phi_{ij}^{ij} \Phi^2 = \chi^2$ .

Hence, it appears that we may interpret Pearson's mean square contingency as a coefficient of dispersion, namely, a measure of the deviation between the mean or expected number a cell should have in the case of independence and the mean or expected number it actually has relative to the mean or expected number a cell should have in the case of independence as a unit of measure summed for all cells.

4. *Multiple and partial contingency.* In the case of three variables, suppose that it is true that

$$(10) \quad F_{ijk} = \phi_{ijk}^i \cdot \phi_{ijk}^j \cdot \phi_{ijk}^k,$$

where  $F_{ijk} = \phi_{ijk}^{ijk} \phi_{ijk}$ .

As before, in the case of independence,

$$(11) \quad \bar{F}_{ijk} = \bar{\phi}_{ijk}^i \cdot \bar{\phi}_{ijk}^j \cdot \bar{\phi}_{ijk}^k.$$

Again, if  $\phi_{ijk}$  is the characteristic function of the observation,

$$(12) \quad \bar{\phi}_{ijk} = p_{ijk}; \quad \bar{\phi}_{ijk}^i = p_{ijk}^i; \quad \bar{\phi}_{ijk}^j = p_{ijk}^j; \quad \bar{\phi}_{ijk}^k = p_{ijk}^k; \quad \bar{\phi}_{ijk}^{ijk} = p_{ijk}^{ijk} = 1.$$

From (10), (11), and (12), in the case of independence, we

find that

$$(13) \quad p_{ijk} = p_{ijR}^i \cdot p_{ijR}^j \cdot p_{ijk}^k,$$

and in the case of dependence, we have

$$(14) \quad p_{ijk} = M(\phi_{ijR}^i \cdot \phi_{ijR}^j \cdot \phi_{ijk}^k) \neq p_{ijR}^i p_{ijR}^j p_{ijk}^k.$$

The quantity  $(p_{ijk} - p_{ijR}^i \cdot p_{ijR}^j \cdot p_{ijk}^k)$  represents the departure between the mean value  $\phi_{ijR}^i$  has and that which it should have in the case of independence.

We now consider the square of the departure relative to  $p_{ijk}^i \cdot p_{ijk}^j \cdot p_{ijk}^k$ , namely,

$$\psi_{ijk}^2 = \frac{(p_{ijk} - p_{ijR}^i \cdot p_{ijR}^j \cdot p_{ijk}^k)^2}{p_{ijk}^i \cdot p_{ijk}^j \cdot p_{ijk}^k}.$$

For all cases, we have

$$(15) \quad \Phi^2 = (\psi_{ijk}^2)^{ijk}$$

which we call the *mean square multiple contingency* in the case of three variables or attributes.

In general, in case we have  $n$  attributes:

$$(16) \quad \psi_{i_1 i_2 \dots i_n}^2 = \frac{(p_{i_1 i_2 \dots i_n} - p_{i_1 i_2 \dots i_n}^1 - \dots - p_{i_1 i_2 \dots i_n}^n)^2}{p_{i_1 i_2 \dots i_n}^1 \dots p_{i_1 i_2 \dots i_n}^n},$$

and for all cases:

$$(17) \quad \Phi^2 = (\psi_{i_1 i_2 \dots i_n}^2)^{i_1 i_2 \dots i_n},$$

which we call the *mean square multiple contingency* in the case of  $n$  attributes.

Let us again consider the case of three attributes. We may write

$$\Phi^2 = \left\{ (\psi_{ij}^2)_{jk}^i \right\}^k = \left\{ (\psi_{ik}^2)_{ij}^i \right\}^j = \left\{ (\psi_{jk}^2)_{ij}^i \right\}^j.$$

For a given  $k$ ,

$$(18) \quad \Phi_k^2 = (\psi_{ij}^2)_{jk}^i$$

is the *partial mean square contingency* between two attributes for an assigned third attribute.

If  $\Phi_k^2 = 0$  for every  $k$  ( $k = 1, 2, 3, \dots$ ), then

$$\Phi^2 = (\Phi_k^2)^k = 0.$$

Similarly, if  $\Phi_i^2$  and  $\Phi_j^2$  are zero for every  $i$  and every  $j$ , respectively, then

$$\Phi^2 = (\Phi_i^2)^i = 0, \text{ and}$$

$$\Phi^2 = (\Phi_j^2)^j = 0.$$

We have thus proved the theorem, namely,

*Theorem 1:* The necessary and sufficient condition for the three attributes to be independent is that

$$(19) \quad \left\{ \begin{array}{l} \Phi^2 = (\Phi_k^2)^k = 0, \text{ or} \\ \Phi^2 = (\Phi_i^2)^i = 0, \text{ or} \\ \Phi^2 = (\Phi_j^2)^j = 0. \end{array} \right.$$

It is fairly easy to see that in the case of  $n$  attributes, we have

$$\Phi^2 = \left\{ (\psi_{i_1 i_2 \dots i_n}^2)_{i_1 i_2 \dots i_n}^{i_1 i_2 \dots i_n} \right\}^{i_1 i_2 \dots i_n}$$

For a given set  $i_3, i_4, \dots, i_n$

$$(20) \quad \bar{\Phi}_{i_3 i_4 \dots i_n}^2 = (\psi_{i_1 i_2}^2)_{i_3 i_4 \dots i_n}^{i_1 i_2}$$

where  $\bar{\Phi}_{i_3 i_4 \dots i_n}$  is the *partial mean square contingency* between two attributes for an assigned set of  $(n-2)$  attributes.

If  $\bar{\Phi}_{i_3 \dots i_4} = 0$  for any pair  $i_1, i_2$ , and for every associated set  $i_3, i_4, \dots, i_n$ , then

$$\bar{\Phi}^2 = \left( \bar{\Phi}_{i_3 i_4 \dots i_n}^2 \right)_{i_3 \dots i_n}^{i_1 i_2} = 0.$$

Hence, we have the

*Theorem 2:* The necessary and sufficient condition for complete independence in the case of  $n$  attributes is that for every pair  $i_1, i_2$ , it is true that

$$(21) \quad \bar{\Phi}^2 = \left( \bar{\Phi}_{i_3 i_4 \dots i_n}^2 \right)_{i_3 i_4 \dots i_n}^{i_1 i_2} = 0.$$

Again, it is fairly easy to see that in general different values assigned to the set  $i_3, i_4, \dots, i_n$  will result in corresponding different values for  $\bar{\Phi}_{i_3 i_4 \dots i_n}^2$ . Hence, if  $\omega \bar{\Phi}_{i_3 i_4 \dots i_n}^2$  is the weighted arithmetic mean of these different values where the respective weights are the relative numbers of individuals in each sub-set, then we say that

$$\omega \bar{\Phi}_{i_3 i_4 \dots i_n}^2$$

is the *partial mean square measure of contingency*.

5. *Mean square dependence.* Rietz<sup>9</sup> invented games of chance which give a meaning to correlation in pure chance. The writer believes it important at least formally to propose a measure of

<sup>9</sup> Rietz, H. L., "Urn schemata as a basis for the development of correlation theory," *Annals of mathematics*, Vol. 21, 1919-20, pp. 306-322.

dependence based upon a probability schemata. As before, let the attributes be  $X$  and  $Y$ .

Let us assume that

$$\begin{aligned} F_{ij} &= F(\phi_{ij}^i, \phi_{ij}^j, i, j). \text{ Then,} \\ \bar{F}_{ij} &= \bar{F}(\phi_{ij}^i, \phi_{ij}^j, i, j), \text{ whence,} \end{aligned}$$

$$p_{ij} = \bar{F}_{ij},$$

where  $\bar{F}_{ij}$  is the mean value of  $F(\phi_{ij}^i, \phi_{ij}^j, i, j)$  and  $p_{ij}$  is the mean value of  $F_{ij}$ .

The quantity  $(p_{ij} - \bar{F}_{ij})$  represents the departure from dependence for the particular  $F(\phi_{ij}^i, \phi_{ij}^j, i, j)$  under discussion. We now form the quantity  $D_{ij}$  defined as

$$(22) \quad D_{ij}^2 = \frac{(p_{ij} - \bar{F}_{ij})^2}{\bar{F}_{ij}},$$

which is the square of the departure relative to  $\bar{F}_{ij}$ .

For all cases, we have

$$(23) \quad \delta^2 = (D_{ij}^2)^{ij},$$

which we call the *mean square dependence*.

Our concept of dependence may be extended to cases of more than two attributes and measures of multiple as well as partial dependence may be obtained in an analogous fashion. It thus appears that we have, at least formally, a general criterion for dependence and an approach to a general criterion which may serve as a *measure of goodness of fit*.

We also note that in every contingency table the events designated by the  $p_{ij}$  or  $\bar{F}_{ij}$  are *mutually exclusive* for every  $i$  and  $j$ .

6. *A measure of connection.* We here propose to idealize Gini's measure of connection which has been fully discussed by

the writer elsewhere.<sup>10</sup> Gini's measure of connection is of interest and importance since one of his special indices of connection is Pearson's correlation ratio and one of his special indices of concordance is Pearson's correlation coefficient. These facts are established in my paper referred to above.

As before, let  $\phi_{ij}^{\delta}$  represent the number of individuals having the group value  $X_j$  of  $X$  and  $Y_i$  of  $Y$  in case we have the two attributes  $X$  and  $Y$ . The total number of individuals having the group value  $Y_i$  of  $Y$  is  $\phi_{ij}^{\delta}$  and the total number of individuals having the group value  $X_j$  of  $X$  is  $\phi_{ij}^{\delta}$ . The total number of individuals is  $\phi_{ij}^{\delta}$ . The frequencies of  $Y$  are distributed according to a set of "partial" groups which correspond to the respective modalities of  $X$ . If all the "partial" groups are similar to the "total" group of frequencies of  $Y$ , then the distribution of modalities of  $Y$  is independent of the modalities of  $X$  and  $Y$  is not connected with  $X$ . In other words,  $Y$  is not dependent upon  $X$  but is independent of  $X$  in the probability sense. Again, if at least one of the "partial" groups is not similar to the "total" group of frequencies of  $Y$ , then the distribution of modalities of  $Y$  is dependent on the modalities of  $X$  and  $Y$  is connected with  $X$ . In other words,  $Y$  is dependent on  $X$  and is not independent of  $X$  in the probability sense.

We now multiply the frequencies of each "partial" group by a number  $w_j$  such that the total frequency of each "partial" group is the same as the number of cases examined. For a given cell, the frequency is then  $w_j \phi_{ij}^{\delta}$  and the total frequency of this "partial" group is then  $w_j \phi_{ij}^{\delta} = \phi_{ij}^{\delta}$ .

Let us now consider the quantity  $G_{ij}^{\delta}$  defined by

$$G_{ij}^{\delta} = \phi_{ij}^{\delta} - w_j \phi_{ij}^{\delta}.$$

The mean value of  $\phi_{ij}^{\delta}$  is  $\bar{\phi}_{ij}^{\delta}$  and the mean value of  $w_j \phi_{ij}^{\delta}$  is

<sup>10</sup> Weida, F. M., "On various conceptions of correlation," *Annals of Mathematics*, Vol. 29, No. 3, July 1928, pp. 276-312.

$p_{ij}^{\cdot}$ . If  $M_{ij}$  is the mean value of  $G_{ij}$  then

$$(24) \quad M_{ij} = p_{ij}^{\cdot} - p_{ij}^{\cdot}.$$

We now consider a quantity  $d_j$  defined by

$$(25) \quad d_j = (|M_{ij}|)^i,$$

which is Gini's *simple index of dissimilarity* and may be regarded as the sum of the absolute values of a set of mean values.

We now consider the quantity  $\phi_{ij}^i d_j$ . The mean value of  $\phi_{ij}^i d_j$  is  $p_{ij}^i d_j$ .

For all cases, the mean value  $I_{YX}$  is given by

$$(26) \quad I_{YX} = (p_{ij}^i d_j)^j,$$

which is Gini's *measure of connection of Y on X*. Thus, Gini's measure of connection may be regarded as the mean value of a set of sums of absolute values of mean values. An analagous discussion holds for  $I_{XY}$  which is Gini's measure of a connection of X on Y.

It is fairly easy to see that the process may be extended to derive measures of multiple, partial and complete connection. This the writer intends to accomplish at a future date.

7. *Conclusion.* It is believed that we have shown that the theory of contingency, dependence and connection may be based upon a definition of probability that includes all forms of probability. Fluctuations in random sampling appear to be neglected in such a treatment, however the experiments may be carried out with the probability schemata in case we desire the inclusion of fluctuations in random sampling.



# NOTE ON KOSHAL'S METHOD OF IMPROVING THE PARAMETERS OF CURVES BY THE USE OF THE METHOD OF MAXIMUM LIKELIHOOD

By

R. J. MYERS

It has been shown by R. A. Fisher<sup>(1)</sup> that the most efficient parameters for Pearsonian curves may be found by the method of maximum likelihood. In applying this method we maximize the quantity

$$(1) \quad L = \sum n_k \log p_k$$

by varying the parameters of the curve;  $n_k$  denotes the observed frequency of the  $k^{\text{th}}$  class, and  $p_k$  is the probability of an observation falling in this class as determined from the curve and is thus a function of the parameters. Thus, in maximizing  $L$ ,  $p_k$  varies as the parameters are varied, but  $n_k$  remains constant throughout since it is fixed by the given data.

Usually it is impossible to obtain a solution to the maximum likelihood equation so that some method of approximation must be used. R. S. Koshal<sup>(2)</sup> has devised a very ingenious method of approximation, which can be summarized briefly as follows. Values of  $L$  are obtained first by varying only one parameter at a time, and then by varying two parameters at the same time. When only one parameter is varied, two values of  $L$  are computed for each parameter, whereas in the case of two parameters being varied, only one value of  $L$  is computed for each combination of parameters. Thus,  $2n + nC_2 + 1$  or  $\frac{1}{2}(n+1)(n+2)$  values of  $L$  would be needed for  $n$  parameters. With these  $L$ 's the constants of  $n$  simultaneous equations involving the  $n$  corrections to the  $n$  parameters can be determined, and then the corrections themselves can readily be obtained.

In applying this method a number of interesting results were

obtained. The data used was the same as used by Koshal<sup>(2)</sup> because in checking through his work there were found several serious numerical errors, especially in the computation of  $\beta$ . This gave a poor fit so that the method of maximum likelihood had more opportunity for improvement than if there had been no error. These data are distributed according to a Type 1 distribution, whose general equation is

$$(2) \quad y = y_0 (x - \alpha)^{m_1} (\beta - x)^{m_2}$$

The values of the parameters as obtained from the moments are

$$\begin{aligned} \alpha &= .33461 \\ \beta &= 16.9885 \\ m_1 &= .69753 \\ m_2 &= 4.93202. \end{aligned}$$

The most convenient sizes of the increments for the parameters were chosen, namely .1 for  $\alpha$ ,  $m_1$ , and  $m_2$  and 1.0 for  $\beta$ .

In the case of the  $L$ 's in which only one parameter is varied, Koshal selected the two  $L$ 's to be computed for a particular parameter in the following manner: it should be remembered that  $L_{0000}$ , the value for the unaltered parameters, has already been computed. As an illustration let us consider the  $L$ 's computed for variations of  $\alpha$ . The criterion set up was that  $L_{\bar{x}+1\ 000}$  should be greater than either  $L_{x\ 000}$  or  $L_{\bar{x}+2\ 000}$ , where  $x$  may be  $-2$ ,  $-1$ , or  $0$ . This criterion is justified by the common sense reasoning that the maximum likelihood solution will then lie somewhere between  $L_{x\ 000}$  and  $L_{\bar{x}+2\ 000}$ . However, in the case of the  $L$ 's in which two parameters are varied, Koshal merely selected the combination of the increments at random. Thus, for the  $L$  for  $\alpha$  and  $\beta$ , Koshal computed  $L_{1100}$ . In carrying out my computations I thought it best to use the same criterion on the  $L$ 's in which two parameters were varied, as was used on the  $L$ 's in which only one parameter was varied. For example, I gave various values to  $x$  and  $y$  so that a number of values of  $L_{xy00}$

were obtained. The largest of these was used in the determination of the constants as explained before. It was not necessary to give all values to  $x$  and  $y$  because a good many combinations could be discarded by inspection. For example, if  $L_{1100}$  was greater than  $L_{1000}$ , it obviously was not necessary to calculate  $L_{1-100}$ .

The above process was repeated for the other  $L$ 's, and the constants were then determined. From these the corrections to the parameters were obtained; these corrections gave new parameters as follows:

$$\begin{aligned}\gamma &= .38399 \\ \beta &= 16.5020 \\ m_1 &= .72547 \\ m_2 &= 4.80853.\end{aligned}$$

The frequency distribution obtained from these parameters was quite a bit better than the original one as judged by both the  $\chi^2$ 's test and its likelihood. However, it is important to note that two of the double increment  $L$ 's used in obtaining the constants were greater than the  $L$  obtained from the new parameters. This would seem to show that better results could be gotten by judicious guessing than by using this method of approximation. Another fact illustrating the roughness of approximation is that the values of the constants when computed from other of the double increment  $L$ 's vary by as much as 30% from those previously used. Naturally with different values of the constants, different values for the corrections to the parameters would be obtained. Several combinations of different values of the constants were tried, and a few of the resulting frequency distributions gave higher  $L$ 's than the ones obtained previously, although there were none higher than the two subsidiary  $L$ 's previously mentioned. It is not unlikely that a combination of constants might be found so as to yield a higher  $L$  than either of the latter two, but there would have to be a considerable amount of manipulation in order to find this combination.

Another disadvantage of this method is the fact that a great deal of time is required to apply it. Approximately sixty hours were required to carry the calculations for the Type 1 curve.

Another interesting fact was brought out when the method of Pearson and Pairman<sup>(3)</sup> for correcting the moments for grouping was applied to the original data. The frequency distribution obtained was far better than any previously obtained as shown by the fact that the  $L$  for this distribution was highest of all;  $\chi^2$  for this distribution was 4.64. The time required to apply this method was considerably less than needed for Koshal's method.

Since writing this paper my attention has been directed to the recent article in the Journal (Vol. XCIII, Part II, 1934, p. 331) by W. P. Elderton and G. H. Hansmann. In this paper the writers used the same data as Koshal and fit these data by an ingenious method due to Elderton<sup>(4)</sup>. It is interesting to note that the  $\chi^2$  of the distribution obtained by Elderton and Hansmann is practically the same as that obtained when the method of Pearson and Pairman was used. Elderton and Hansmann also came to the conclusion that Koshal's method required more labor to bring about the same results as other methods.

#### BIBLIOGRAPHY

1. Fisher, R. A. "On the Mathematical Foundations of Theoretical Statistics." *Phil. Trans.*, A, vol. 222, pp. 309-368.
2. Koshal, R. S. "Application of the Method of Maximum Likelihood to the Improvement of Curves Fitted by the Method of Moments." *Jour. Royal Stat. Soc.*, vol. XCVI, pp. 303-313.
3. Pairman, Eleanor and Pearson, Karl. "On Corrections for the Moment-Coefficients of Limited Range Distributions when there are Finite or Infinite Ordinates and any Slopes at the Terminals of the Range." *Biometrika*, vol. 12, pp. 231-258.
4. Elderton, W. P. "Frequency Curves and Correlation," pp. 121-122, 2nd edition.

# THE ADEQUACY OF "STUDENT'S" CRITERION OF DEVIATIONS IN SMALL SAMPLE MEANS\*

By

ALAN E. TRELOAR AND MARIAN A. WILDER  
*Biometric Laboratory, University of Minnesota*

## INTRODUCTION

The origin of the movement toward precise evaluation of probabilities based on the statistics of small samples would generally be located by practical statisticians in the work of "Student" (1908). The problem he considered is of such importance, not only from the historical aspect, but also from a consideration of the elements of statistical interpretation, that we wish to return to an analysis of the adequacy of his solution. "Student" was concerned with the problem of determining the significance to be attached to the deviation of the mean,  $\bar{x}$ , of a small sample from a probable (or possible) supply† mean,  $m$ , when the dispersal of variates in the supply is unknown. The solution he suggested was based upon derivation of the probability integral of the quantity

$$(1) \quad z = \frac{\bar{x} - m}{s},$$

where  $s$  is the standard deviation of the sample. He found the distribution of  $z$  to be given by the equation,

$$(2) \quad df = k_z (1 + z^2)^{-\frac{N}{2}} dz.$$

In 1915, Fisher indicated that "Student's" partly intuitive derivation was sound, and in 1925 he returned to a more complete exposition of the accuracy of the solution, at the same time widely

---

\* Presented in part before a Joint Session of the Econometric Society and Section K of the American Association for the Advancement of Science, Boston, Dec. 30, 1933.

† Following Wickseil (e.g. *Biometrika* 25, p. 121), we shall use the term "supply" in place of "population."

extending its application. Fisher at that time changed the variable to  $t = z \sqrt{n}$ , where  $n$  is the number of "degrees of freedom" involved in estimating  $\sigma$  (the supply standard deviation) from  $s$ . "Student" (1925) cooperated in this extension by preparing tables of the probability integral of  $t$ , using  $n$  in place of  $N$  as the parameter. Since the integrals are of essentially identical curves, and  $z$  will prove somewhat more adaptable in the present study, we will conduct the discussion of the problem in terms of  $z$ . All conclusions reached will apply with equal validity, of course, when  $t$  is used in place of  $z$ .

"Student" illustrated the usefulness of his  $z$  distribution by considering the  $x$  values as a set of differences (between experimental and control pairs, say), thus logically making  $m$  equal to zero. He then found the probability that the resulting  $z$  would be exceeded solely through random sampling errors. Although it is not by any means clear from "Student's" original memoir that he so intended, the custom has grown of considering this probability as that which might be expected for the deviation of  $\bar{x}$  from  $m$  if a knowledge of  $\sigma$  were available. Is such a transfer of the probability really acceptable? The usefulness of the  $z$  (or  $t$ ) test depends entirely on the answer to this question.

### SIGNIFICANT DEVIATIONS

In a supply of variates,  $x$ , whose frequency distribution accords with the "normal" curve and whose total frequency approaches infinity, let the mean be  $m$  and the standard deviation  $\sigma$ . Assume a large number of samples, each of total frequency  $N$ , to be drawn independently and at random from this supply. Let the mean and standard deviation of each sample be designated as  $\bar{x}$  and  $s$  respectively. Then the probability that values of  $\bar{x}$  will deviate from  $m$  by more than a certain amount may be determined exactly from the "normal" integral. Letting

(3)

$$y = \frac{\bar{x} - m}{\sigma}$$

the distribution of  $y$  will be given by the equation

$$(4) \quad df = k_y e^{-\frac{N}{2} y^2} dy,$$

a "normal" curve with mean at zero and standard deviation of  $N^{-\frac{1}{2}}$ . Values of  $y$  exceeding  $1.96/\sqrt{N}$  will arise but 5 times in 100, and this value would be known therefore as the "5% level of significance." For  $N$  equal to 5, this level is .8765.

Let a single sample of 5 individuals, not known to be drawn from the above supply, be made available. It may be desired to test whether the mean,  $\bar{x}'$ , of this sample differs sufficiently from  $m$  to warrant the assumption, on the basis of the mean value alone, that the sample has not been drawn from the above supply. If  $(\bar{x}' - m)/\sigma$  should exceed .8765, those depending on a 5% "level of significance" would decide that the sample is significantly different in the respect tested. However,  $y$  will exceed this level 5 times in 100. It must therefore be expected that up to 5% of samples like that designated by the prime above which are investigated by this procedure will be erroneously segregated as "differing significantly."

This maximum error of 5% is acceptable to most workers for two reasons:

(i) Some such error must be accepted in order to have a basis for differentiation, and 5% or less (generally less) erroneous segregation is sufficiently small to be regarded by many as an acceptable proportion of error;

(ii) The cases erroneously segregated in this manner are the most rational ones to be subjected to the error, since they deviate from  $m$  by the greatest amount.

In practical statistical problems wherein the significance of the deviation of a mean is to be tested, it is usually impossible to apply the above reasoning because of lack of precise knowledge of the value of  $\sigma$ . "Student's" test aimed to meet this deficiency by finding the integral of  $z$  already defined (equations 1 and 2).

Applying the probability integral of this variable, he reached his conclusions about the significance of  $z$  in the same way as has been indicated for the variable  $y$ .

### THE CORRELATION BETWEEN $\bar{x}$ AND $s$ .

In analyzing the adequacy of the procedure suggested by "Student," it seems fruitful to consider the correlation of  $\bar{x}$  and  $s$ . Defining the latter in its original sense,

$$(5) \quad s = \sqrt{\sum (x - \bar{x})^2 / N},$$

"Student" (unknowingly justifying Helmert's previous work) concluded the distribution of  $s$  is given by

$$(6) \quad df = \kappa_s e^{-\frac{N}{2} \left(\frac{s}{\sigma}\right)^2} s^{N-2} ds.$$

This most important equation has not received the discussion it deserves. Tables of the probability integral of  $v$ , where

$$(7) \quad v = s/\sigma$$

would also be most helpful in small sample analysis, if for no other reason than to show the wide variation which must be expected in  $s$  for small values of  $N$ . An appreciation of this variation is much more pertinent to the adequate solution of the problem analyzed by "Student" than appears to have been realized. We accordingly include here the 2½% points\* in  $v$  for a few values of  $N$  small.

2½% points for $v$		
$N$	lower	upper
2	.02	1.58
3	.13	1.57
4	.22	1.53
5	.31	1.49

\* By 2½% points we mean those points at which the ordinate truncates a tail whose area is 2½% of the total area of the curve.



It will be seen from these figures that, for  $N$  equal to 5,  $s$  will vary over the relatively very wide range of  $.31\sigma$  to  $1.49\sigma$  even when only the central 95% of cases are considered. Inasmuch as there is no correlation between  $(\bar{x} - m)$  and  $s$  when sampling is made from a "normal" supply, the values to be expected for  $z$  in those samples where  $(\bar{x} - m)$  is the same must vary widely solely through the influence of variation in  $s$ .

Expressing  $(\bar{x} - m)$  and  $s$  in terms of  $\sigma$  as the unit of measurement, the simultaneous distribution we wish to analyze will become that of  $y$  and  $v$ . Since these variables are wholly independent (see Fisher, 1925), their simultaneous distribution will be given by the product of their separate probabilities, yielding

$$(8) \quad df = k_{y,v} e^{-\frac{N}{2}y^2} e^{-\frac{N}{2}v^2} v^{N-2} dy dv.$$

This surface is graphically portrayed in Figure 1 for the case when  $N$  equals 5. The few contours given are sufficient to indicate the general character of the distribution of frequency. Projection of the frequencies onto the two margins gives the univariate distributions drawn in the Figure.

If  $B$  and  $B'$  be taken as the  $2\frac{1}{2}\%$  points for the  $y$  distribution, then lines through them drawn perpendicular to the  $y$  axis will cut off in the extreme zones of the surface and in the tails of the  $y$  distribution those samples whose means deviate sufficiently from  $m$  to permit their segregation according to a "5% level of significance."

Since

$$z = y/v,$$

the samples segregated by the 5% level in applying the  $z$  test must be bounded on one side (in each direction) by radial lines traversing this surface and passing through the point  $(y=0, v=0)$ . Let  $b$  be the value of the  $2\frac{1}{2}\%$  point for the  $z$  distribution. Then the cotangent of the angle of incidence to the  $y$  axis will in each case equal  $b$ , i.e. 1.3882 when  $N$  equals 5.

All samples given by points in the shaded areas,  $E$  and  $F$  (Fig-

ure 1), would be considered significantly deviating with respect to  $\bar{x}$  according to customary interpretation of the  $z$  test. Those samples in the shaded areas,  $F$  and  $G$ , would be segregated by the  $y$  test. Only those samples in the cross-shaded regions,  $F$ , would be selected by both tests. For the situation under discussion, wherein the sampling is actually made from the one supply, no samples really deviate in  $\bar{x}$  from  $m$  by an amount not logically to be ascribed to random sampling effects. For reasons given earlier in this discussion, however, the  $y$  segregates are all rationally made. Only the  $z$  segregates in the double-shaded area  $F$  may be designated as rational on the grounds given. Those in the single-shaded area  $E$  are irrationally selected; the segregation has been made because  $s$  is small, not because  $(\bar{x} - m)$  is large.

#### THE CORRELATION BETWEEN $\bar{x}$ AND $z$ .

An analogous geometric view may be presented by considering the correlation surface for  $y$  and  $z$ . To obtain the simultaneous distribution of these variables, the substitutions

$$v = y/z,$$

$$dv = -\frac{y}{z^2} dz,$$

may be made in equation (8), yielding:

$$(9) \quad df = k_{y,z} e^{-\frac{N}{2} y^2} y^{N-1} e^{-\frac{N}{2} \left(\frac{y}{z}\right)^2} z^{-N} dy \cdot dz$$

In slightly different form, Pearson (1931a) has given this expression and derived from it the equations for the correlation, regression and scedasticity of the surface in terms of  $N$ . He demonstrated that, although regression is rectilinear and  $r_{\bar{x}z}$  is very high, the distribution of  $z$  for constant  $\bar{x}$  is characterized by "excessive leptotosis and extreme skewness" for  $N$  small, with gradual approach to "normality" as  $N$  increases. Also, there is marked heteroscedasticity of these arrays.

It is a simple matter to truncate the  $(y, z)$  surface into volumes of frequency corresponding to the probability of occurrence of given deviates in  $y$  or  $z$ . This is graphically portrayed in Figure 2, where the surface is approximately represented for  $N = 5$  and the planes of truncation,  $BCD$  and  $bCd$ , correspond to the 2.5% points,  $B$  and  $b$  respectively, for each variable. Since the frequency surface is radially symmetrical about the point  $(y = 0, z = 0)$ , only one quadrant need be lettered. 2.5% of the area of the "normal"  $y$  distribution lies in the minor segment bounded by the ordinate  $AB$ , and 2.5% of the "leptokurtic"  $z$  distribution lies in the minor segment bounded by the ordinate  $ab$ . Also, 2.5% of the total frequency of the correlation surface lies in the two minor volumes truncated by the vertical planes passing through  $AB$  and  $ab$  respectively. Only that proportion of frequency lying beyond *both* planes, i.e. in the area  $bCd$ , exceeds the given level for both variables simultaneously.

The corresponding frequency volumes in Figures 1 and 2 representing segregations by the  $y$  and  $z$  tests are as follows:

## FIGURE 1

Zone  $E$ Zone  $F$ Zone  $G$ 

## FIGURE 2

Zone  $dCD$ Zone  $bCD$ Zone  $BCb$ 

That the corresponding zones should not have the same relative areas in the two figures is in accordance with expectation, since the densities of frequency must vary widely within the zones and in different manners from one zone to another. Interpretation of the degrees of rational and irrational segregation by the  $z$  test must depend upon evaluation of the integrals defining the respective frequency volumes.

## EVALUATION OF INTEGRALS

For the  $(y, v)$  surface, the frequency over each double-shaded zone  $E$  will be given by the expression

$$(10) \quad \Delta f_i = \kappa_{y,v} \int_B^\infty e^{-\frac{N}{2} y^2} dy \int_0^{by} e^{-\frac{N}{2} v^2} v^{N-2} dv.$$

For the  $(y, z)$  surface, the corresponding frequency over the area,  $bCD$ , will be given by the expression

$$(11) \quad \Delta f_z = \kappa_{y,z} \int_B^\infty e^{-\frac{N}{2} y^2} y^{N-1} dy \int_b^\infty e^{-\frac{N}{2} \left(\frac{y}{z}\right)^2} z^{-N} dz.$$

The constants,  $\kappa_{y,v}$  and  $\kappa_{y,z}$ , prove to be identical in magnitude, and we shall therefore give the evaluation of the latter only.

Integrating from zero to infinity in both directions, one secures half the total frequency since the distribution appears equally and solely in the two quadrants of positive product.

$$\frac{1}{2 \kappa_{y,z}} = \int_0^\infty e^{-\frac{N}{2} y^2} y^{N-1} dy \int_0^\infty e^{-\frac{N}{2} \left(\frac{y}{z}\right)^2} z^{-N} dz$$

But

$$\begin{aligned} \int_0^\infty e^{-\frac{N}{2} \left(\frac{y}{z}\right)^2} z^{-N} dz &= \frac{2^{\frac{N-3}{2}}}{N^{\frac{N-1}{2}} y^{N-1}} \int_0^\infty u^{\frac{N-3}{2}} e^{-u} du \\ &= \frac{2^{\frac{N-3}{2}} \sqrt{\frac{N-1}{2}}}{N^{\frac{N-1}{2}} y^{N-1}}. \end{aligned}$$

Therefore

$$\frac{1}{2 \kappa_{y,z}} = \frac{2^{\frac{N-3}{2}} \sqrt{\frac{N-1}{2}}}{N^{\frac{N-1}{2}}} \int_0^\infty e^{-\frac{N}{2} y^2} dy$$

$$= \frac{2^{\frac{N-1}{2}} \sqrt{\left(\frac{N-1}{2}\right)} \pi^{1/2}}{N^{\frac{N}{2}}}$$

and

$$(12) \quad k_{y,z} = \frac{N^{\frac{N}{2}}}{2^{\frac{N-2}{2}} \sqrt{\left(\frac{N-1}{2}\right)} \pi^{1/2}}$$

It is pertinent to prove now that  $\Delta f_1$  equals  $\Delta f_2$ .

Letting  $w = \frac{N}{2} v^2 = \frac{N}{2} \left(\frac{y}{z}\right)^2,$

then  $dw = N v dv = -\frac{N y^2}{z^3} dz.$

Substituting in (10), we have,

$$\int_0^{\frac{y}{z}} e^{-\frac{N}{2} v^2} v^{N-2} dv = \frac{2^{\frac{N-3}{2}}}{N^{\frac{N-1}{2}}} \int_0^{\frac{N y^2}{2 z^2}} e^{-w} w^{\frac{N-3}{2}} dw$$

Substituting in (11), we have,

$$\int_0^{\infty} e^{-\frac{N}{2} \left(\frac{y}{z}\right)^2} z^{-N} dz = \frac{2^{\frac{N-3}{2}}}{N^{\frac{N-1}{2}} y^{N-1}} \int_0^{\frac{N y^2}{2 z^2}} e^{-w} w^{\frac{N-3}{2}} dw.$$

Thus

$$(13) \quad \Delta f_1 = \frac{N^{\frac{1}{2}}}{\sqrt{\left(\frac{N-1}{2}\right)} \pi^{1/2}} \int_0^{\infty} e^{-\frac{N}{2} \frac{y^2}{z^2}} dy \int_0^{\frac{N y^2}{2 z^2}} e^{-w} w^{\frac{N-3}{2}} dw = \Delta f_2.$$

Noting that  $B$  equals  $1.96/\sqrt{N}$ , it would seem logical to conclude from the general form of equation (13) that  $\Delta f$  approaches a limit of .025 as  $N$  increases. We have not yet succeeded in proving this explicitly.

Numerical evaluation of the double integral for  $\Delta f$  presents difficulties. These may be overcome by applying a succession of reduction formulas to the series of single integrals in powers of  $y^2$  obtained from the integration with respect to  $w$ . For example, when  $N = 5$ ,  $B = 0.8765$ ,  $b = 1.3882$ , and

$$\begin{aligned}\Delta f &= \frac{\sqrt{5}}{\sqrt{2\pi}} \int_B^\infty e^{-\frac{5}{2}y^2} dy \int_0^{\frac{5y^2}{2b^2}} e^{-w} w dw \\&= \frac{\sqrt{5}}{\sqrt{2\pi}} \left[ \int_B^\infty e^{-\frac{5}{2}y^2} dy - \frac{5}{2b^2} \int_B^\infty e^{-\frac{5}{2}y^2(1+\frac{1}{b^2})} y^2 dy - \int_B^\infty e^{-\frac{5}{2}y^2(1+\frac{1}{b^2})} dy \right] \\&= .025 - \frac{\sqrt{5} \cdot B \cdot e^{-\frac{5B^2}{2}(1+\frac{1}{b^2})}}{2\sqrt{2\pi} (b^2+1)} - \frac{\sqrt{5}}{\sqrt{2\pi}} \left[ 1 + \frac{1}{2(b^2+1)} \right] \int_B^\infty e^{-\frac{5}{2}y^2(1+\frac{1}{b^2})} dy \\&= .025 - .0072 - \frac{b \cdot (2b^2+3)}{2(b^2+1)^{3/2}\sqrt{2\pi}} \int_{\frac{B}{b}}^\infty e^{-\frac{w^2}{2}} dw \\&= .0178 - .0074 = .0104\end{aligned}$$

Values for the frequency volumes  $\Delta f$  (corresponding to the area  $bCD$  in Figure 2) are given as column (4) of Table I for the chosen values of  $N$ . The differences between these values and .025 provide the magnitudes of the frequency volumes corresponding to  $BCb$  and  $dCD$ . The latter volumes, which are necessarily equal, are given in column (5) of the same table. In columns (6) and (7) the values in columns (4) and (5) respectively are expressed as percentages of the limiting value, .025.

We have not succeeded as yet in expressing any of these proportional frequencies as simple equations in terms of  $N$  only. In Figure 3, however, a graph of the relationship is plotted, based on the data of Table I. The vertical scale on the left gives the proportional frequency beyond the two planes passing through  $C$ . By following the dotted lines to the scale on the right vertical margin, the percentage error ( $100 dCD/.025$ ) with which we are concerned may be read off directly.

TABLE I

Data for evaluation of volumes truncated by the planes passing through  $C$  (Fig. 1), for different sizes of sample, where  $C$  corresponds to the .025 points of  $y$  and  $z$ .

(1)	(2)	(3)	(4)	(5)	(6)	(7)
$N$	$B$	$b$	Volumes*		Volumes* as % of .025	
			$bCD$	$BCb=dCL$	$bCD$	$BCb=dCD$
3	1.1316	3.042	.0064	.0186	25.6	74.4
5	.8765	1.388	.0104	.0146	41.6	58.4
7	.7408	.999	.0126	.0124	50.4	49.6
9	.6533	.815	.0142	.0108	56.8	43.2
11	.5910	.705	.0151	.0099	60.4	39.6
13	.5436	.629	.0160	.0090	64.0	36.0
15	.5061	.573	.0166	.0084	66.4	33.6
17	.4754	.530	.0171	.0079	68.4	31.6
19	.4497	.495	.0176	.0074	70.4	29.6
21	.4277	.466	.0179	.0071	71.6	28.4
25	.3920	.421	.0185	.0065	74.0	26.0
29	.3640	.387	.0190	.0060	76.0	24.0
99	.1970	.201	.0216	.0034	86.4	13.6

## PRACTICAL TESTS

In order to test the accuracy of the above deductions when applied to a supply which is grouped into fairly fine categories, two sampling studies were made. Samples of 5 individuals each were drawn in both cases. The first study dealt with a much used supply of two anthropometric measures which conform fairly well to the "normal" curve in their distributions. The second study

\* Volumes (of frequency) follow the notation of Figure 2.

used as a supply a *theoretical* "normal" bivariate frequency surface, seriated into classes. These studies will be referred to as Series I and II.

*Series I.* From the table provided by MacDonell (1902) on the associated variation of stature (to the nearest inch) and length of the left middle finger (to the nearest millimeter) in 3000 British criminals, the measurements were transferred to 3000 numbered Denison metal-rim tags from which the cords had been removed. After thorough checking and mixing of these circular disks, samples of 5 tags each were drawn at random until the supply was exhausted. Unfortunately, three of these samples were erroneously returned to a receiving box before being copied, and the records of 597 samples only are available. For these, the statistics  $y$  and  $z$  were calculated for each variable, and frequency surfaces for joint occurrence of  $y$  and  $z$  were prepared in which the statistics for stature and finger length were first considered separately, then combined. After calculating the correlation coefficient, the frequencies of the opposite quadrants were added so as to provide the seriation without regard to the signs of  $y$  and  $z$ . The actual number of cases falling beyond the planes of truncation corresponding to the 2.5% points were then counted and the proportional frequencies tabled.

*Series II.* From the tables of the probability integral of the "normal" correlation surface prepared by Lee and others (see Pearson, 1931*b*) a correlation table of total frequency of 1000 approximately was prepared for the case where the correlation is .5, using  $.3\sigma$  as the unit of classification in both directions. Modification of the fractional frequencies to the nearest whole number yielded a table in which  $N$  equalled 998,  $r$  equalled .5003 and the two standard deviations equalled .9914 (Sheppard's correction applied). Samples of 5 were drawn by working systematically through the tables of random numbers provided by Tippett (1927), 2043 samples being so secured. These samples were treated as in the case of Series I.



The actual correlation surfaces secured for the joint occurrence of  $y$  with  $v$  and  $y$  with  $z$  may be illustrated by scatter diagrams prepared from the data of Series II. These are given as Figures 4 and 5, the variates in the latter case being considered without regard to sign. Both conform very well indeed to the theoretical contour diagrams presented earlier (Figures 1 and 2).

For the correlation between  $y$  and  $z$  (signs *not* ignored), Pearson (1931*a*) has determined theoretically that, for  $N$  equal to 5,  $r$  should equal  $+.8862$ . We find the following results for our two series:

Variable	Series I	Series II
1	.8744	.8797
2	.7883	.8869
1 + 2	.8270	.8832

For Series II the agreement with theory is splendid. The wider deviations from the theoretical value in Series I are probably due, in part, to the less perfectly "normal" nature of the supply distributions.

The inadequacy of the correlation coefficient as a descriptive measure of such a "non-normal" surface as that for  $y$  and  $z$  will be apparent at once from an inspection of figure 5. Discordance of the two variables increases rapidly as their values increase to such an extent that, for  $N$  equal to 5, values of  $z$  beyond the customary level of significance provide exceedingly poor bases of prognostication concerning the true significance of the deviation in the mean, despite the fairly high value of the correlation coefficient.

In Table II the frequencies beyond the chosen levels of significance for  $y$  and  $z$ , separately and jointly, are given for both series. The empirical frequencies are given in Roman type in the whole numbers, and as proportions in parentheses. The theoretical values are given in italics in the last column for comparison. The agreement is very good in every case, the deviation of observed values

from the theoretical being well within the range of error assignable to random sampling effects.

TABLE II

Comparison of actual and theoretical frequencies beyond the given levels of significance in the practical tests

Series	I	II	Theoretical
Total frequency	1194 (1)	4096 (1)	1
Frequency beyond 5% level for			
(a) $y$ alone	56(.0469)	206(.0504)	.05
(b) $z$ alone	59(.0494)	191(.0467)	.05
(c) $y$ and $z$ together	22(.0184)	79(.0193)	.0208
Maximum inefficiency of $z$ test	62.7%	58.6%	58.4%

## SUMMARY

"Student's" distribution has been very widely used in the analysis of small samples in order to determine the probability that the deviation of a mean is ascribable to errors of random sampling. Most workers appear to have lost sight of the fact that the distribution is that of a ratio, in which both the numerator and denominator must be expected to vary independently. It is quite erroneous to ascribe the probability of such a ratio to the value taken by the numerator alone.

The rationality of segregation according to any given "level of significance" using "Student's" distribution may be analyzed by considering the joint distributions due to errors of sampling in the means, standard deviations, and the ratio of these two for samples of any given size,  $N$ . Theoretical evaluation of the percentage of irrationally segregated samples is given herein for the odd values of  $N$  from 3 to 29 and for  $N = 99$ , using the 5% level of significance. This percentage falls in a curvilinear manner as  $N$  increases, a few values being 75% for  $N = 3$ , 58% for  $N = 5$ , 33% for  $N = 15$ , and 14% for  $N = 99$ . The so-called "large" samples, then, are open to a considerable error of this kind. These

results have been verified by two extensive sampling tests for the case where  $N = 5$ .

Results such as those given herein stress again the dangers attendant upon the drawing of deductions of practical importance from a single sample of small size. When only a single sample is available it is certainly desirable that the statistical analysis should depend not merely upon most likely estimates of needed parameters, but also upon those of less probability which might readily be true and which guard against the erroneous segregation of possibly insignificant deviations.

#### LITERATURE CITED

FISHER, R. A.

1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* 10: 507-521.

FISHER, R. A.

1925. Application of "Student's" distribution. *Metron* 5: 2-32.

MACDONELL, W. R.

1902. On criminal anthropometry. *Biometrika* 1: 177-227.

"STUDENT"

1908. The probable error of a mean. *Biometrika* 6: 1-25.

"STUDENT"

1915. Tables for estimating the probability that the mean of a unique sample of observations lies between  $-\infty$  and any given distance of the mean of the population from which the sample is drawn. *Biometrika* 11: 414-417.

PEARSON, KARL

- 1931a. Some properties of "Student's"  $z$ : Correlation, regression and scedasticity of  $z$  with the mean and standard deviation of the sample. *Biometrika* 23: 1-9.

PEARSON, KARL

- 1931b. Tables for statisticians and biometricians. Part II. Cambridge University Press, England. pp. ccl + 262.

TIPPETT, L. H. C.

1927. Random sampling numbers. Cambridge University Press, England. pp. viii + 26.

#### ACKNOWLEDGMENT

Our thanks are most heartily extended to Professor Dunham Jackson of the University of Minnesota for suggesting the analysis of the  $y, v$  surface as an alternative method of elucidating the

problem, which was first explored in terms of the  $y, z$  association; also to Professor Harold Hotelling of Columbia University for helpful criticisms of an earlier draft of this paper. Very material assistance has also been given by a grant-in-aid from the Rockefeller Foundation through the Graduate School Research Fund of the University of Minnesota.

FIGURE 1

Theoretical frequency surface for  $y$  and  $v$ , separately and jointly, for  $N = 5$ .

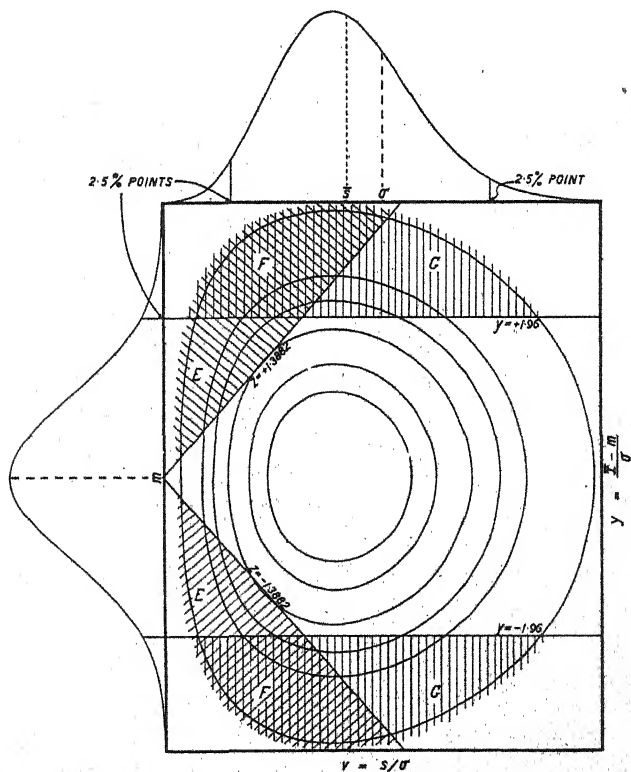


FIGURE 2

Theoretical frequency distributions of  $y$  and  $x$ , separately and jointly, for  $N=5$ . (Contours for the joint distribution are approximate only and the intervals between them do not correspond to the same increment of frequency.)

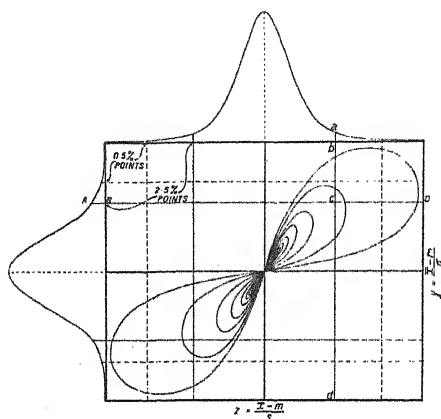


FIGURE 3

Curve to illustrate the increase in correct segregation of means by the  $x$  test as  $N$  increases.

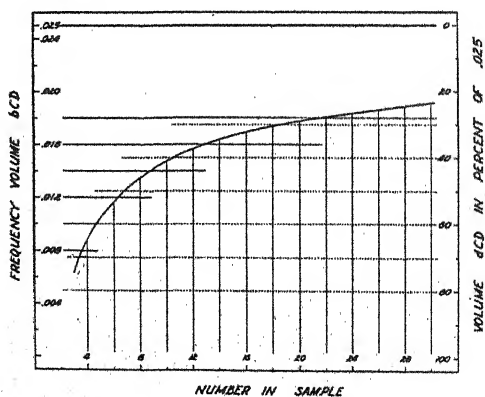


FIGURE 4

Frequency surface for the joint occurrence of  $y$  and  $v$  as secured in Series II.

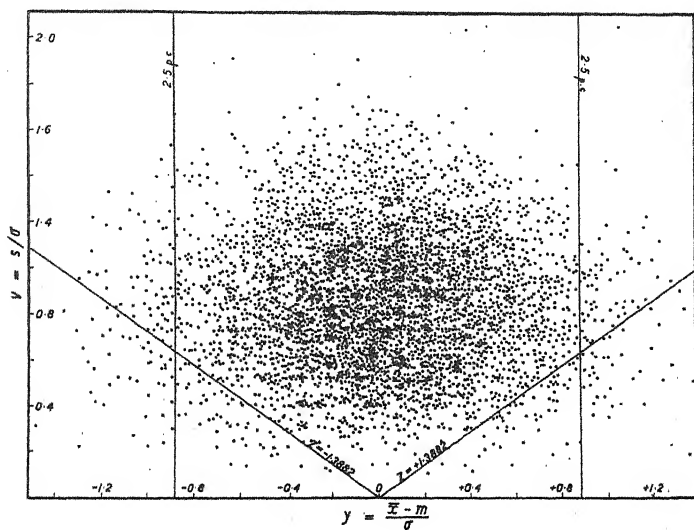
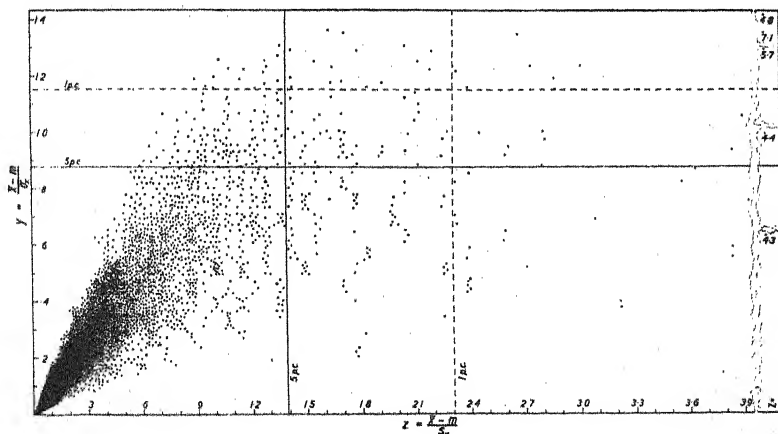


FIGURE 5

Frequency surface for the joint occurrence of  $y$  and  $z$  as secured in Series II.



(All Rights reserved)

# BIOMETRIKA. Vol. XXVI, Parts III and IV

## CONTENTS

	PAGES
I. The Wilkinson Head of Oliver Cromwell in relation to Portraits, Busts, life and Death Masks. By KARL PEARSON and G. M. MORANT. With 106 Plates . . .	269—378
II. Contribution à l'Étude de la Théorie de la Corrélation. Par CARLOS E. DIEULEFAIT .	379—403
III. The Use of Confidence or Fiducial Limits illustrated in the Case of the Binomial. By C. J. CLOPPER and EGON S. PEARSON. With five Diagrams in the Text . . .	404—413
IV. The Roumanian Silhouette. By MARIOARA PERTIA and Others. With two Plates, Map, Diagram, two Figures in Text and two Contours in Pocket . . .	414—424
V. On a New Method of Determining "Goodness of Fit" By KARL PEARSON . . .	425—442
VI. A Statistical Study of the <i>Daucus carota</i> L. (Second Article.) By WILLIAM DOWELL BATES. With eleven Figures in the Text . . .	443—468

### MISCELLANEA:

Review of Paul Harzer's <i>Tabellen für alle statistische Zwecke in Wissenschaft und Praxis</i> . By F. GARWOOD . . .	469—470
---	---------

The publication of a paper in *Biometrika* marks that in the Editors' opinion it contains either in method or material something of interest to Biometricians. But the Editors desire it to be distinctly understood that such publication does not mark assent to the arguments used or to the conclusions drawn in the paper.

A volume of *Biometrika* containing about 400 pages, with plates and tables, is issued annually.

Papers for publication and books and offprints for notices should be sent to Dr KARL PEARSON, University College, London. It is a condition of publication in *Biometrika* that the paper shall not already have been issued elsewhere, and will not be reprinted without leave of the Editors. It is very desirable that a copy of all measurements made, not necessarily for publication, should accompany each manuscript. In all cases the papers themselves should contain not only the calculated constants, but the distributions from which they have been deduced. Diagrams and drawings should be sent in a state suitable for direct photographic reproduction, and if on decimal paper it should be blue ruled, and the lettering only pencilled.

Papers will be accepted in French, Italian or German. In the last case the manuscript should be in Roman not German characters.

Contributors receive 25 copies of their papers free. Joint authors 15 copies each. Fifty additional copies may be had on payment of 17/- per sheet of eight pages, or part of a sheet of eight pages, with an extra charge for Plates; these should be ordered when the final proofs are returned.

The subscription price, payable in advance, is 45s. net per volume: single issues 54s. net (including postage) for Great Britain, and 54s. net abroad (including packing and postage). Owing to the scarcity of early volumes, the following rates must now be charged for complete sets. Vols. I—XXV, including XX<sup>P</sup>: Inland, bound in buckram £106, in wrappers £98 net; abroad £124. 15s. in buckram, £114. 15s. in wrappers. Recent volumes may still be obtained at wrapper prices. Standard buckram cases with Darwin block, price 8s. 6d. + 6d. postage per volume. Index to Vols. I to V, 2s. net. Index to Vols. I to XV, 7s. 6d. net. Cheques must be made payable to Dr Karl Pearson and sent to The Secretary, *Biometrika Office*, Zoological Laboratory, University College, London, W.C. 1, to whom all orders for series and single copies should be addressed. All cheques must be properly stamped and should be crossed "*Biometrika Account*." No foreign cheques can be accepted unless they are drawn in sterling, properly stamped, and payable at a London agency.

**Indian Agricultural Research Institute (Pusa)**  
**LIBRARY, NEW DELHI-110012**

This book can be issued on or before .....

Return Date	Return Date